Implementation of a Voice Recognition System in the Nasa Yuwe Language Based on Convolutional Neural Networks

Implementación de un sistema de reconocimiento de voz en el lenguaje Nasa Yuwe basado en Redes Neuronales Convolusionales

Implementação de um sistema de reconhecimento de voz na linguagem Nasa Yuwe baseado em Redes Neurais Convolucionais

Julio Enrique Muñoz Burbano[1]
Pablo Emilio Jojoa Gómez[2]
Fausto Miguel Castro Caicedo[3]

Received: October 2nd, 2022
Accepted: December 15th, 2022
Available: January 22th, 2023

How to cite this article:
J.E. Muñoz Burbano, P.E. Jojoa Gómez, F.M. Castro Caicedo "Implementation of a Voice Recognition System in the Nasa Yuwe Language Based on Convolutional Neural Networks," *Revista Ingeniería Solidaria,* vol. 19, no. 1, 2023.
doi: https://doi.org/10.16925/2357-6014.2023.01.01

Research article. https://doi.org/10.16925/2357-6014.2023.01.01

[1] Estudiante de Maestría Universidad del Cauca
Email: je.munoz@unicauca.edu.co
ORCID: https://orcid.org/0000-0003-2797-8503
CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0002018648

[2] Docente, Universidad del Cauca
Email: pjojoa@unicauca.edu.co
ORCID:https://orcid.org/0000-0002-8461-4063
CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000310840

[3] Docente, Universidad Nacional Abierta y a Distancia
Email: miguel.castro@unad.edu.co
ORCID:https://orcid.org/0000-0002-3017-6328
CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001600363

Abstract

Introduction: This paper presents the Implementation of an algorithm for voice recognition in the Nasa Yuwe language based on Convolutional Neural Networks (CNN), developed at the Universidad del Cauca in the year 2022.

Problem: The Nasa Yuwe language is phonetically rich, as it has 32 vowels and 34 consonants, which leads to confusion in pronunciation and therefore difficulties in recognizing voice patterns.

Objective: To implement a speech recognition algorithm for the Nasa Yuwe language supported in CNN.

Methodology: The preprocessing of the audio signals was carried out to subsequently obtain the characteristics through the scalograms of the Mel coefficients. Finally, an architecture of the CNN is proposed for the classification process.

Results: A DataSet is built from the scalograms of the voice patterns, and the CNN training process is carried out.

Conclusion: The implementation of a Voice Recognition System based on CNN provides low margins of error in the word classification process of the Nasa Yuwe language.

Originality: The proposed voice recognition system is the first and only one of its kind that has been carried out so far, with the purpose of collaborating in the process of teaching, preserving and learning the Nasa Yuwe language.

Limitations: It is necessary to increase the number of voice patterns provided by native speakers, and there is a need to implement other technological tools that allow for the conservation and dissemination of the Nasa Yuwe language.

Keywords: VRS (Voice Recognition System), Nasa Yuwe Language, Mel coefficients, CNN (Convolutional Neural Network), Machine Learning

Resumen

Introducción: Este artículo presenta la Implementación de un algoritmo para el reconocimiento de voz en el lenguaje Nasa Yuwe basado en Redes Neuronales Convolusionales (RNC), desarrollado en la Universidad del Cauca en el año 2022.

Problema: La riqueza fonética del lenguaje Nasa Yuwe es grande, al poseer 32 vocales, y 34 consonantes, lo que lleva a confusiones en la pronunciación y por lo tanto a dificultades en el reconocimiento de patrones de voz.

Objetivo: Implementar un algoritmo de reconocimiento de voz, para el lenguaje Nasa Yuwe soportado en RNC.

Metodología: Se realizó el preprocesamiento de las señales de audio para posteriormente obtener las características por medio de los escalogramas de los coeficientes de Mel. Finalmente se propone una arquitectura de la RNC para el proceso de clasificación.

Resultados: Se construye un DataSet a partir de los escalograma de los patrones de voz, y se realiza el proceso de entrenamiento de la RNC.

Conclusión: La implementación de un SRV basado RNC, proporciona bajos márgenes de error en el proceso de clasificación de palabras del lenguaje Nasa Yuwe.

Originalidad: El sistema de reconocimiento de voz planteado es el primero y único en su clase que se ha realizado hasta el momento, con el propósito de colaborar en el proceso de enseñanza, conservación y aprendizaje del lenguaje Nasa Yuwe.

Limitaciones: Se requiere aumentar el número de patrones de voz aportados por hablantes nativos, y se plantea la necesidad de implementar otras herramientas tecnologías que permitan la conservación y difusión del lenguaje Nasa Yuwe.

Palabras clave: SRV (Sistema de Reconocimiento de Voz), Lenguaje Nasa Yuwe, coeficientes de Mel, RNC (Redes Neuronales Convolucionales), Machine Learning.

Resumo

Introdução: Este artigo apresenta a Implementação de um algoritmo para reconhecimento de voz na linguagem Nasa Yuwe baseado em Redes Neurais Convolucionais (RNC), desenvolvido na Universidade de Cauca em 2022.

Problema: A riqueza fonética da língua Nasa Yuwe é grande, pois possui 32 vogais e 34 consoantes, o que gera confusão na pronúncia e, consequentemente, dificuldades no reconhecimento dos padrões vocais.

Objetivo: Implementar um algoritmo de reconhecimento de fala para a linguagem Nasa Yuwe suportada pelo RNC.

Metodologia: Foi realizado o pré-processamento dos sinais de áudio para posterior obtenção das características através dos escalogramas dos coeficientes de Mel. Finalmente, uma arquitetura RNC é proposta para o processo de classificação.

Resultados: Um DataSet é construído a partir do escalograma dos padrões de voz, e o processo de treinamento do RNC é realizado.

Conclusão: A implementação de um SRV baseado em RNC fornece baixas margens de erro no processo de classificação de palavras da língua Nasa Yuwe.

Originalidade: O sistema de reconhecimento de voz proposto é o primeiro e único do gênero realizado até o momento, com o objetivo de colaborar no processo de ensino, conservação e aprendizagem da língua Nasa Yuwe.

Limitações: É necessário aumentar o número de padrões de voz fornecidos por falantes nativos, e levanta-se a necessidade de implementar outras ferramentas tecnológicas que permitam a conservação e disseminação da língua Nasa Yuwe.

Palavras-chave: SRV (Sistema de Reconhecimento de Fala), Linguagem Nasa Yuwe, Coeficientes de Mel, RNC (Redes Neurais Convolucionais), Aprendizado de Máquina.

1. INTRODUCTION

A Voice Recognition System (VRS) represents the set of techniques and algorithms used by a computational tool to identify and transform a voice pattern, in such a way that it allows human-machine interaction, with the purpose of solving a certain need [1]. VRS have gained special relevance, especially due to their versatility and functionality in applications and different uses such as medicine, robotics and home automation technologies, among others, making the devices that handle this type of interface increasingly precise and easier to handle [2] [3] [4]. VRS models typically have the following stages: Acquisition of the voice signal, pre-processing of the signal,

recognition and classification of the patterns of the voice signal [5]. To achieve a VRS with reliable margins of error, it is necessary to use techniques for extracting the characteristics of the voice patterns[6], among the most used are: 1-) the scalograms obtained from the Mel coefficients, the MFCC (Mel Frequency Cepstral Coefficients), which are associated with the neurological perception of sound, and in most cases are used to identify the speaker and not the spoken word [7]; 2-) the characteristics provided by the application of the Wavelet transform, which allow a wide range of possibilities given the number of existing families, although the wavelet transforms are usually sensitive to volume variations in the audio samples [8]; 3-) the Fourier transform that provides information on speech patterns in the frequency domain, but does not provide information on its temporal variations [9]. An advantage of VRSs is that they can work with a large amount of vocabulary without difficulty, and the processing time is low, allowing the user to satisfactorily assess the performance of the VRS [10].

Among the commercial uses of VRS, we find virtual assistants such as Siri (Apple) [11], Cortana (Microsoft) [12], Google Assistant and Google Now (Google) [13], Alexa (Amazon) [14] and Bixby (Samsung)_[15] [16]. Also, among other basic applications of the VRS are: Villamil's proposal in 2005 [17], for the development of voice recognition tools such as isolated words and connected digits; recognition of short sentences [18], and different applications in medicine [19]. As can be seen, Speech Recognition has become a transversal and interdisciplinary research area applicable in many fields; however, most of the applications are for widely used languages such as English, so to apply it in a language or specific language such as Nasa Yuwe, it is necessary to make the adaptations and investigations of each case.

Nasa Yuwe is the language spoken by the Nasa people, who are concentrated in seven departments of the Republic of Colombia: Cauca, Huila, Tolima, Valle del Cauca, Meta, Caquetá and Putumayo; being in Cauca where the largest population occurs [20]. It is currently a language in danger of extinction due to the different interactions of the indigenous community with other communities [21], various cultural, social, geographical and even historical factors, to the point that there are communities where Nasa Yuwe has become a second language. Therefore, it is necessary to support its revitalization from different perspectives, such as tools that favor processes of pronunciation and language teaching; it is there where voice recognition systems become important, since they could support mechanisms to practice and improve bilingualism processes within the Nasa Community [22].

In the literature there is no virtual assistant or voice recognition system for Nasa Yuwe, therefore, this document presents the experience of designing and building a

VRS for the Nasa-Yuwe language, which makes use of convolutional neural networks (CNN), which have led to notable advances in the efficiency of speech recognition engines. The system proposed here becomes the first speech recognition system used for the recognition and classification of speech patterns in Nasa Yuwe.

The document layout is as follows: In Section 2, a brief description of the context is made and the methodological process followed is presented. Section 3 illustrates the creation of the DataSet, the proposed architecture for the CNN, the training process and comparison of the performance of the CNN. Finally, in Section 4, the discussions and conclusions are presented.

1.1 RELATED WORKS.

The different tools and techniques provided by Machine Learning and Deep Learning allow for a wide variety of applications in various fields of research, where the use of these tools can be appreciated for processes of classification, regression and prediction models.

The use of Machine Learning tools can be found in the work carried out by Contreras, Caro and Morales in 2022, in which a review of Ensemble methods for predicting student academic performance is carried out. In the study, two periods of time between 2016 to 2018 and 2019 to 2021 are related. For each of these periods of time, the different techniques used in prediction models, the size of the samples and the precision corresponding to the best model are evidenced [23].

An application of Deep Learning is exposed in the work presented by Balakrishna, Gopi and Solanki, who in 2022 carried out a comparative analysis of some deep learning models in the detection of cyberbullying in social networks. The models compared were the long-term memory model (LSTM), bidirectional long-term memory (BI-LSTM), recurrent neural networks (RNN), bidirectional recurrent neural networks (BI-RNN), closed recurrent unit (GRU) and Bidirectional Closed Recurrent Unit (BI-GRU). The models are applied to the Twitter social network and their performance is evaluated through some metrics such as Accuracy, F1, Recall, Precision. The work proposes a methodology for obtaining and preprocessing the data to later be implemented in the Deep Learning models. The corresponding analysis of results shows an improved accuracy of 90.4% [24].

In 2020, Calp and Kose propose the estimation of burned areas in forest fires through the use of Artificial Neural Networks. In the proposed work, they describe the preprocessing of the data and their normalization with the purpose of building a relevant DataSet. Later, they propose an architecture of the Artificial Neural Network

e-ISSN 2357-6014 / Vol. 19, no. 1 / january-april 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

and its corresponding training in the MATLAB® software environment. The Artificial Neural Network is applied to different regions to estimate the burned areas, obtaining very low margins of error and reliable results in the training processes [25].

In recent years, some work has been carried out related to the study and classification of voice patterns for native languages, that is, languages that are still spoken in some regions or countries, and that retain their own characteristics and particular distinctive elements.

Within these studies, we find the work carried out in 2013 by Suuny, Peter and Poulose, who propose a system for automatic voice recognition for the Malayalam language, which seeks the classification of ten classes corresponding to the voice patterns of the pronunciation of numbers from zero to ten. The extraction of characteristics was carried out by means of the Wavelet transform method. Later, they carried out the training of an Artificial Neural Network with an accuracy of 89% [26].

In the case of South America, the research work presented by Maldonado, Villalba and Pinto-Roa, in 2016, is taken as a reference, in which the development of an automatic voice recognition system for Guaraní speech is appreciated, this work applies techniques modeling of Hidden Markov Models with probability-based approaches, in order to obtain a relevant result in the speech pattern classification process [27].

We also found the research presented by Aimituma and Churata in 2019, who presented a speech-to-text converter for the Quechua language. The extraction of characteristics was carried out by obtaining the MFCC and, for the classification process, a hybrid model was trained that is made up of HMM and a Deep Neural Network. At the end of the research, they obtained a speech recognition accuracy of 59.20% [28].

2. METHODOLOGY

2.1 The Nasa Yuwe languaje.

The language of the Nasa indigenous community is known as Nasa Yuwe, which has phonetically rich and has also been classified as an independent language, since it is not similar to any other in the world. The Nasa Yuwe has 32 vowels and 34 consonants, as can be seen in Table 1.

e-ISSN 2357-6014 / Vol. 19, no. 1 / january-april 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

Table 1. Unified alphabet Nasa Yuwe.[29]

Voweles				
Oral: i e a u		**Nasal**: ĩ ẽ ã ũ		
Interrupted Orals: i' e' a' u'		**Interrupted Nasals**: ĩ' ẽ' ã' ũ'		
Aspirate Oral: ih eh ah uh		**Aspirate Nasal**: ĩh ẽh ãh ũh		
Large oral: ii ee aa uu		**Large Nasal**: ĩĩ ẽẽ ãã ũũ		
Consonants	**bilabial**	**Alveolar**	**palatal**	**Velar**
Básics	p	T	ç	K
Aspirated	ph	Th	çh	Kh
Palatalized	px	Tx	çx	Kx
Palatalized Aspitrated	pxh	Txh	çxh	Kxh
Prenasal	b	D	z	G
Palatalized -Prenasal	bx	Dx	zx	Gx
Nasal	m	N		
Palatalized -Nasal		Nx		
Fricative		S		J
Palatalized- Fricative	fx	Sx		Jx
Lateral		L		
Palatalized -Lateral		Lx		
Approximants	w		y	
Palatalized - Approximants	vx			

Source: own work

As can be seen, it is not an easy language to pronounce, which can have an impact on the fact that a small change in pronunciation changes the meaning of the words pronounced [30]. For example, the case of the word *Ate* whose translation into Spanish is "clean", and the word A'te which translates to "moon"; or the word *Eç* that translates "document" and the word E'ç whose translation is "emerald" [31]. The apostrophe " ' " after the vowel represents a glottization in pronunciation [32].

2.2 Convolutional Network Net.

A Convolutional Neural Network (CNN) is defined as a set of processing layers, see Figure 1, where each layer of the CNN can contain three main variables: inputs, weights and outputs, where the output of a layer becomes the input of the next layer. This process is sequential and not linear [33] [34].

e-ISSN 2357-6014 / Vol. 19, no. 1 / january-april 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

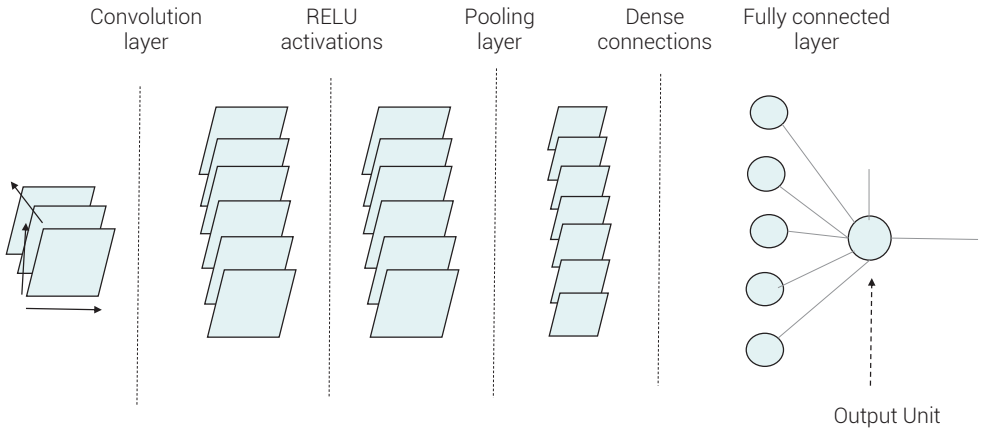

Figure 1. Basic structure of an CNN.
Source: own work adapted from [35]

The convolutional layer consists of a group of kernels or filters that can learn, these filters scan different sets of pixels to determine the RGB composition of each one and calculate the products between the filter input and the input data at any other position. The volume of the input data is analyzed, a two-dimensional activation map is obtained, these maps are the response of the filter to each spatial position. After training, the CNN will learn from the characteristics extracted from the filters and will recognize certain patterns.

In the activation layer, the RELU (Rectified Linear Units) formula is applied to each of the neurons. It is possible to consider it as a function that has a constant gradient, while the total input is positive and has a slope with zero value for other inputs.

$$\mathrm{RELU}(x)=\begin{cases} x\leq 0 \rightarrow x=0 \\ x>0 \longrightarrow x=0 \end{cases}. \tag{35}$$

On the other hand, the purpose of the Pooling layer is to reduce the resolution of the feature maps. This implies that the units of this layer will serve as generalizations about the features of the lower convolution layer, and thanks to this process, the generalizations will once again be spatially localized in frequency and will be invariable to minimal variations in their corresponding location. This reduction is achieved by applying a pool function to multiple drives in a local region of a size determined by a parameter called pool size. The Pooling algorithm significantly reduces the spatial size of the representation. Specifically, the most used algorithm is MaxPooling, which consists of taking a sample of the most representative neurons before continuing with the following convolutional layers.

The Full-Connected (FC) layer consists of neurons that are fully connected to the neurons of the previous layer. It also analyzes the extracted vector and determines to which class it belongs.

For the optimal functioning of the CNN, it is necessary to generate an image that allows for the analysis of patterns for the VRS and to carry out an organization of the input data for the CNN, which can be done by means of a two-dimensional (2-D) matrix that can be the result of the scalogram, which provides information about the logarithmic energy calculated directly from the Mel frequency spectral coefficients of a voice pattern. For a color image, the values of the different colors, red, green and blue (RGB), can be seen as three different 2-D feature maps. In this case, the CNN executes a small window on the input image in the training and testing phases, in such a way that it can begin its process of learning the functions of the input data regardless of the absolute position of the data within the image matrix.

In general terms, the benefits of CNN for VRS can be summarized in three fundamental properties: locality, shared weight and clustering. Locality implies that the most salient features can be computed locally from cleaner parts of the spectrum and only a smaller number of features are affected by noise; Shared weighting allows each weight to be learned from multiple frequency bands on the input instead of learning from a single location; Clustering makes it easy for the same feature values computed at different locations to be represented by one value, which in turn leads to minimal differences in the features extracted by the clustering layer [36].

2.3 Mel coefficients

The MFCC are coefficients for speech representation based on human auditory neurological perception. The MFCC show the local characteristics of the voice signal associated to the vocal tract (depending on the instant of analysis), according to the filter-source model. The MFCC are derived from the Fourier transform (FT - Fourier Transform), and the discrete cosine transform (DCT - Discrete Cosine Transform) is also included, but the basic peculiarity is that, in MFCC, the frequency bands are located in a logarithmic scale according to the Mel scale [37].

In general, humans can differentiate tones between 100 Hz and 200 Hz and not tones between 1000 Hz and 1100 Hz, even though the frequency difference between them is 100 Hz. Therefore, it is difficult for human beings to differentiate between higher frequencies and easier to differentiate between lower frequencies. This implies that, although the frequency distance between the two sounds is the same, the perception

e-ISSN 2357-6014 / Vol. 19, no. 1 / january-april 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

of the human being is not the same; which is why the Mel scale is essential for audio machine learning applications, since it imitates the human perception of sound [7].

In the digital processing of voice signals, MFCC are used in several applications, the main ones being: speech recognition, synthesis and coding. All of these applications require the amount of acoustic wave information to be reduced to allow processing by a computational tool. In this sense, the process of extracting features from the voice signal requires that such features should not contain information redundancy, in order to maximize the performance of the systems that use them [38].

2.4. Materials and methods

2.4.1 Data.

11 classes that correspond to different Nasa Yuwe words were taken as the basis for the classification, see Table 2.

Table 2. Nasa Yuwe words

Nasa-Yuwe word	Translation
Ab	Groove
Afx	Clear, transparent, clean
Akh	Cold, flu
Bajx	Warm
Jabx	Surly
Kasx	Basket
Kath	Tight, stick, steep slope
Kazx	Ugly, bad
Khlala	Boil
Lavx	Smooth, slippery
Sap	Toad

Source: own work

Each of the words was pronounced 3 times by native speakers, which were an adult woman, a girl and two adult men. Voice patterns are the basis for feature extraction by means of Mel coefficients.

e-ISSN 2357-6014 / Vol. 19, no. 1 / january-april 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

2.4.2 Proposed voice recognition system and methodology

In the design of the CNN-based VRS, the following stages were considered, see Figure 2:

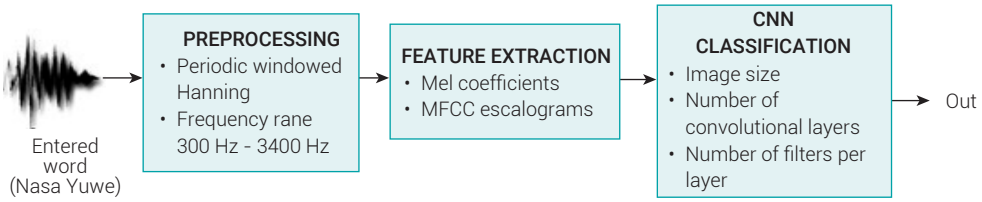

Figure 2. Stages of the proposed VRS.
Source: own work

For the preprocessing of the audio signals, a process of denoising and normalization of the real duration of the sample that varies between 0.3 and 1 second was carried out. Periodic Hanning window filtration was performed to eliminate discontinuities in the samples [39]. The frequency range was established between 300 and 3400 HZ, which guarantees coverage of the frequency ranges of the human voice [40].

The feature extraction was carried out by obtaining the MFCC, and for the scalograms of the MFCC, the number of bands (NB) was varied from 100 (default value in MATLAB®) to 40 with scales of 20 in 20 to form the groups of graphs of the classes. This was done because for large values of NB, significant errors are generated in the graphs and it is necessary to make changes in the lengths of the fast Fourier transforms (FFT), which are typical of the process of obtaining the MFCC, which is not recommended because the length of the FFT would be outside the parameter proposed for Hanning windowing. On the other hand, for small values of NB, the graphs lose detail and the information they provide is minimal.

In the classification stage, we worked with images of the scalograms whose size was 227x227x3 and the original RGB was also maintained.

3. RESULTS

This section presents the creation of the training DataSet, the CNN architecture, the training parameters and the corresponding comparison of the CNN performance with other classification techniques.

e-ISSN 2357-6014 / Vol. 19, no. 1 / january-april 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

3.1 DataSet creation and feature extraction.

It is necessary for the CNN to train with images obtained from the selected features of the audio patterns of words in Nasa Yuwe. Each of the Nasa Yuwe words was pronounced three times by four native speakers: two adult men, an adult woman and a girl, thus having 12 tracks in WAV format for each of the classes to be classified

Subsequently, each graph was labeled with the word in Nasa Yuwe and a number from 1 to 48 according to its order within the set corresponding to each class.

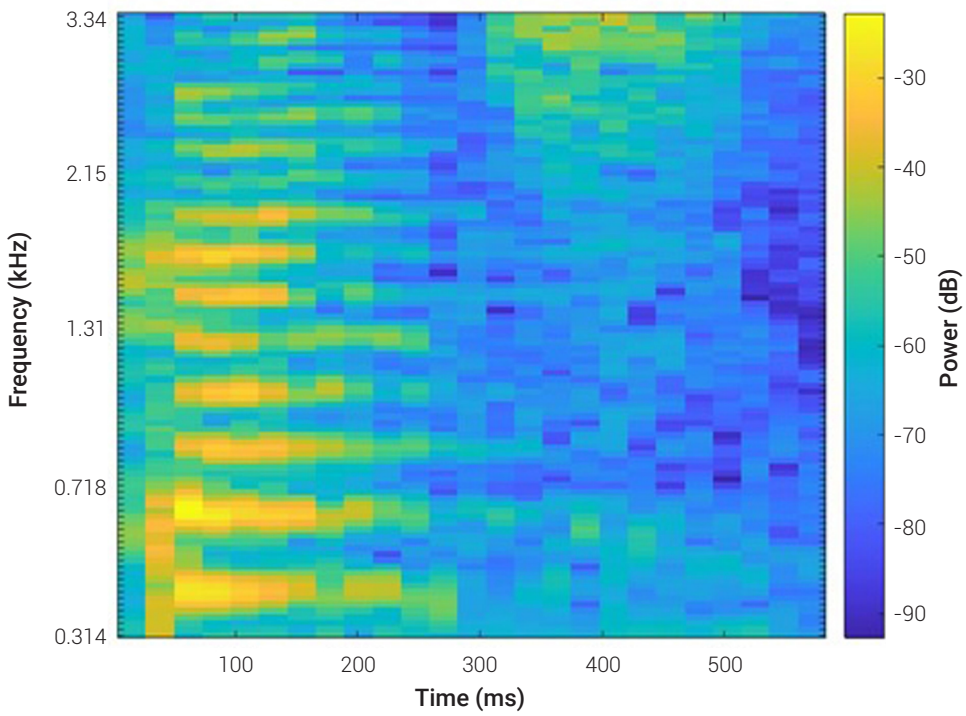

Figure 3. MFCC escalogram for the word KAZX with NB 100. Label kazx8.
Source: own work

e-ISSN 2357-6014 / Vol. 19, no. 1 / january-april 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

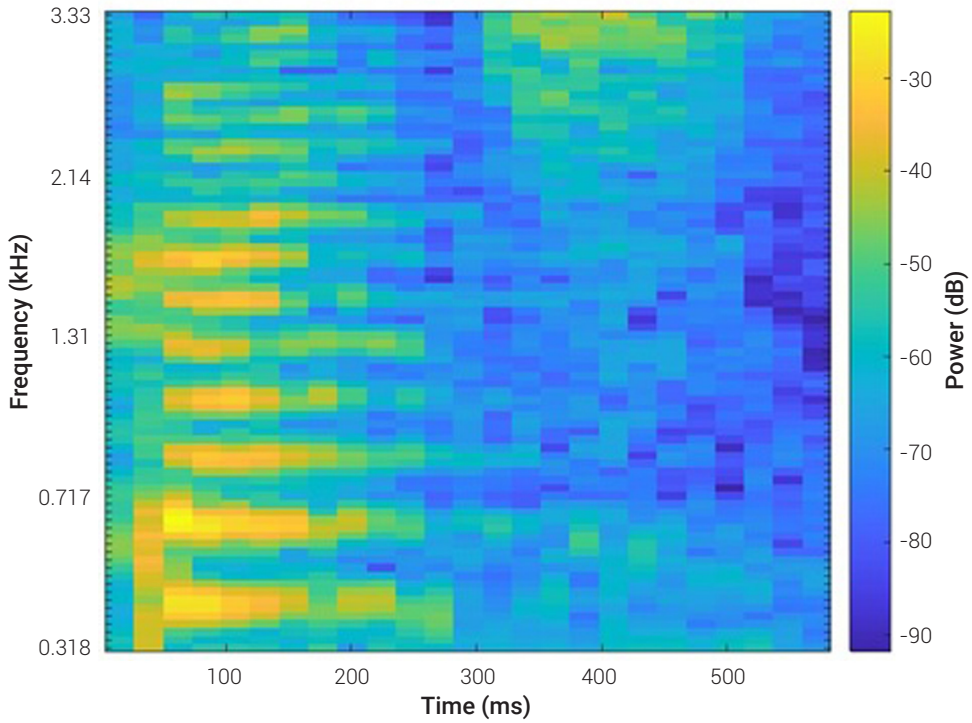

Figure 4. MFCC escalogram for the word KAZX with NB 80. Label kazx20.
Source: own work

3.2 CNN Architecture

The CNN architecture was defined, starting from the input layer where the training DataSet is specified along with an initial set of validation that is part of the initial DataSet; the percentage was 70% for training and 30% for validation.

The training options for the CNN were as follows:

- The training was carried out with sigmoid
- Training tests were performed with the computer's CPU (i7 Processor) and also with the GPU (NVIDIA GFORCE 760 M Card)
- 10 training epochs and a minibach of 5, to compensate for the size of the DataSet.
- The validation process is done every series and a validation frequency of 1 was maintained.

4 convolutional layers were taken into account, a filter size of 3 x 3 was taken as a base, and the number of filters starts from 8 and increases to 16, 32 and 64 in each

e-ISSN 2357-6014 / Vol. 19, no. 1 / january-april 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

following layer. Given the operation of the CNN, the increase in the size of the filters in each layer allows significant details to be obtained from the scalograms to achieve a better classification process. At the end of the convolutional layer, there is the RELU activation layer and the POOLING layer. Finally, there is a FULL-CONNECTED layer with the total number of classes to classify.

With these specifications, the network training was carried out in MATLAB® 2021, obtaining an accuracy of 98.34% for 30 training series and a standard deviation of 1.27%.

For the comparison of the training process of the CNN, a DataSet formed with vectors (Sparse representation) [41] was created, which correspond to the Mel coefficients of the initial scalograms. Subsequently, two Machine Learning techniques were trained, which were the Ensemble technique and the Tree Optimized technique [42] [43] [44].

For the Ensemble technique, the hyperparameters referring to the maximum number of learners, the maximum learning rate and the number of predictors per sample were optimized, and an average accuracy of 88.84 % was obtained, for 30 training series, with a standard deviation of 4.05%. For the Tree optimized technique, whose hyperparameters were optimized to the maximum number of divisions for each class, an average accuracy of 66.05% was obtained, for 30 training series, and a standard deviation of 0.05%.

We can summarize the comparison in the next table:

Table 3. Accuracy results of the different techniques used.

Technique	Accuracy (%)	Standard Deviation (%)
CNN	98.34	1.27
Ensemble	88.40	4.05
Tree Optimized	66.05	0.05

Source: own work

The CNN training process had different results for DataSets created with audios of the selected words spoken by native and non-native speakers, with the preprocessing and feature extraction described in Section 3.

The results of the training process for the DataSet carried out with native speaker audios are shown in Table 4, where metrics such as: Precision, which allows us to measure the quality of the training model in classification tasks, were also obtained; the Recall, which allows us to appreciate the number of patterns that the

training model allows us to identify; and the F1 value, that represents the Precision and the Recall in a single value to compare the combined performance of both [45].

Table 4. Performance Metrics for Native Speaker DataSet

Measure	Value (%)
Accuracy	98.70
Precision	98.70
Recall	99.08
F1	98.98

Source: own work

For the DataSet of non-native speakers, the obtained results are shown in Table 5

Table 5. Performance Metrics for Non-Native Speaker DataSet

Measure	Value (%)
Accuracy	50.64
Precision	59.36
Recall	50.80
F1	54.79

Source: own work

4. DISCUSSIONS AND CONCLUSIONS.

The CNN training, with respect to the classification process of the selected audio patterns, evidenced significant results by reaching an average accuracy of 98.34%, compared to the Machine Learning techniques selected for the corresponding comparison, Ensemble (average accuracy of 88.40%) and Tree Optimized (average accuracy of 66.05%); starting from an adequate preprocessing of the audio signals that implies a periodic Hanning window, which covered a frequency range between 300 Hz and 3400 Hz, which provided a standardization of the initial conditions of the voice patterns. Subsequently, the study carried out from the characteristics obtained through the scalograms of the Mel coefficients, allowed for the creation of a relevant DataSet. Once the DataSet was created, an efficient CNN architecture was designed, which allowed for optimal computational performance by reducing the CNN training time and solving the need to obtain more voice patterns for each class.

The implementation of the VRS based on CNN resulted in low margins of error and adequate metrics (accuracy, precision, Recall and F1) for a Dataset created with native speaker audio patterns. For the DataSet created with audio patterns of non-native speakers, a significant variation of the metric values was observed, which is due to the difficulty that non-native speakers have in pronouncing the words in Nasa Yuwe, with the demands and requirements of the phonetics of the language.

In the research process developed, it was possible to verify that the implementation of a VRS based on Deep Learning tools, such as the CNN, a pioneer in its class, highlighted the need to develop and deepen the creation and use of computational tools that contribute significantly to the process of teaching, conservation and dissemination of the Nasa Yuwe language.

4.1 FUTURE WORKS

In order to strengthen and deepen the research carried out, some suggestions for subsequent work are proposed below, taking into account the results achieved in the process of implementing the voice recognition system for the Nasa Yuwe language.

Improve and expand the database: The problem suggested by the pronunciation and vocal articulation of words in Nasa Yuwe requires having a significant variety of native speakers, of different ages and of different gender, recording audio tracks with the correct pronunciation of some words.

Increase classes for classification: With a more robust database, the classification process of a greater number of Nasa Yuwe words can be carried out, with the algorithm implemented in this research.

Disclosure of the classification algorithm: With the verified classification algorithm, it would be possible to bring its operation to mobile devices and online platforms to facilitate access for native speakers and non-native speakers who wish to practice and learn more about the Nasa Yuwe language.

Implement the algorithm in connected words: With low margins of error in the classification of words, pertinent adjustments could be made to the algorithm and implemented in the recognition of phrases or short sentences (Speech Recognition).

References

[1] K. Barrios, J. Lopez, S. Mendieta, R. Benavides, Y. Saez, “Portal de Revistas Academica UTP,” 2018. doi: https://doi.org/10.33412/rev-ric.v4.0.1827

[2] J. Camargo, Universidad Pontificia Bolivariana, 2010. [Online]. Available: https://www.mendeley.com/catalogue/c2bf0045-f1c5-342d-a896-212cf29b980e/?utm_source=desktop&utm_medium=1.19.8&utm_campaign=open_catalog&userDocumentId=%7Bfceb5c8-bd0c-36c9-bcf6-eef575d78eec%7D. pp 23-28.

[3] M. Atibi, A. Issam, M. Boussaa, A. Bennis, ResearchGate, 2016. doi: http://dx.doi.org/10.1109/CSIT.2016.7549469

[4] O.L. Ramos, D.A. Rojas, L.A. Góngora, "Reconocimiento de patrones de habla usando MFCC y RNA," *Visión electrónica*, vol.10, no. 1, .pp 5-11. doi: https://doi.org/10.14483/22484728.11712

[5] O. Pérez, F. Poceros, A. Villalobos, "DSpace Tesis IPN," 2013. [Online]. Available: https://tesis.ipn.mx/jspui/bitstream/123456789/12309/1/Sistema%20de%20Seguridad%20por%20Reconocimiento%20de%20Voz%20%28Tesis%20de%20Ingenieria%20ESIME%29.pdf. pp 22-27.

[6] J. Pérez, A. Araujo, "Academia," Noviembre 2018. [Online]. Available: https://www.academia.edu/38038688/Aplicaci%C3%B3n_de_una_Red_Neuronal_Convolucional_para_el_Reconocimiento_de_Personas_a_Trav%C3%A9s_de_la_Voz. pp 78-80.

[7] M. Cruz, F. Lozano, C. Higuera, Repositorio Uniandes, 2021. [Online]. Available: https://repositorio.uniandes.edu.co/handle/1992/50650. pp 2-3.

[8] P. Freeman, V. Kashyap, R. Rosner, Q. Lamb, IOPSience, 2002. [Online]. Available: https://iopscience.iop.org/article/10.1086/324017/pdf. pp 187-188.

[9] J. Bernal, P. Gomez, J. Bobadilla, ResearchGate, 2009. [Online]. Available: https://www.researchgate.net/publication/239813705_Una_vision_practica_en_el_uso_de_la_Transformada_de_Fourier_como_herramienta_para_el_analisis_espectral_de_la_voz. pp 79-81.

[10] E. Villca, S. Carmina, DDIGITAL-UMSS, 2020. [Online]. Available: https://ddigital.umss.edu.bo:8080/jspui/handle/123456789/20216.

[11] Apple, appleinsider, [Online]. Available: https://appleinsider.com/inside/siri.

[12] Microsoft, Microsoft, 2022. [Online]. Available: https://support.microsoft.com/es-es/topic/-qu%C3%A9-es-cortana-953e648d-5668-e017-1341-7f26f7d0f825. pp 1

[13] S. Geek, Social Geek, 2022. [Online]. Available: https://socialgeek.co/tech/google-assistant-google-now-te-contamos-diferencias/. pp 1

[14] Amazon, Amazon, 2022. [Online]. Available: https://developer.amazon.com/es-ES/alexa. pp 1

[15] Samsung, Samsung, 2022. [Online]. Available: https://www.samsung.com/co/support/mobile-devices/how-can-i-use-the-bixby-application/. pp 1

[16] Marketing XXI, Marketing XXI, 2018. [Online]. Available: https://www.marketing-xxi.com/voice-search-asistentes-voz-altavoces-inteligentes-seo-sem/asistentes-virtuales-voz. pp 1

[17] I. Villamil, Pontificia Universidad Javeriana de Colombia, 2005. [Online]. Available: https://www.javeriana.edu.co/biblos/tesis/ingenieria/tesis95.pdf.

[18] R. Fatmi, S. Rashad, R. Integlia, Mendeley, 2019. doi: https://dx.doi.org/10.1109/CCWC.2019.8666491

[19] Z. Alkareem, A. Tajudin, M. Al-Betar, A. Abasi, S. Makhadmeh, N. l Salih, ACM Digital Library, 2019. doi: https://dx.doi.org/10.1145/3321289.3321327

[20] Instituto Colombiano de Cultura Hispánica, Geografía Humana de Colombia. Región Andina Central, tom. IV, vol. II, Bogotá, 2008. pp 1

[21] T. Rojas, DOCERO, 2006. [Online]. Available: https://docero.mx/doc/por-los-caminos-de-la-recuperacion-de-la-lengua-paez-4krn88zr31.

[22] Universidad del Cauca, CRIC-PEBIl-Comisión General de Lenguas, Estudio Sociolingüistico Fase preliminar. Base de datos - CRIC 01/2007 Lengua Nasa Yuwe y Namtrik. Popayàn, Cauca, Colombia, CRIC, Popayán - Colombia, 2008. pp 1

[23] L. Contreras, J. Caro, D. Morales, *Ingenieria Solidaria*, 2022. doi: https://doi.org/10.16925/2357-6014.2022.02.01

[24] S. Balakrishna, Y. Gopi, V. Solanky, *Ingenieria Solidaria*, 2022. doi: https://doi.org/10.16925/2357-6014.2022.01.05

[25] H. Calp, U. Kose, *Ingenieria Solidaria*, 2020. doi: https://doi.org/10.16925/2357-6014.2020.03.08

[26] S. Sunny, D. Peter, K. Poulose, ResearchGate, 2013. doi: https://ijret.org/volumes/2013v02/i04/IJRET20130204032.pdf

[27] D. Maldonado, R. P.-R. D. Villalba, Repositorio Institucional de la UNLP, 2016. [Online]. Available: https://sedici.unlp.edu.ar/bitstream/handle/10915/56979/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y.

[28] F. Aimituma y R. Churata, Repositorio Universidad Nacional De San Antonio Abad Del Cusco, 2019. [Online]. Available: https://repositorio.unsaac.edu.pe/bitstream/handle/20.500.12918/4321/253T20190384_TC.pdf?sequence=1&isAllowed=y.

[29] M. Farfán Martínez, T. Rojas Curieux, Zuy Luuçxkwe kwe'kwe'sx ipx kwetuy piyaaka, Cartilla de aprendizaje de nasa yuwe como segunda lengua, Buenos Aires, 2010. pp 1

[30] G. Alvarez, ResearchGate, 2012. [Online]. Available: https://www.researchgate.net/publication/262753111_A_classifier_model_for_detecting_pronunciation_errors_regarding_the_Nasa_Yuwe_language%27s_32_vowels.

[31] Cabildos Nasa., Scribd, 2005. [Online]. Available: https://es.scribd.com/doc/143624645/Diccionario-Nasa-Yuwe-Castellano. pp 19-81.

[32] T. Rojas, Utexas, 2001. [Online]. Available: https://lanic.utexas.edu/project/etext/llilas/cilla/rojas.html. Pagina web. pp 1

[33] W. Rivas, B. Mazón, ResearchGate, 2018. [Online]. Available: https://www.researchgate.net/profile/Bertha-Mazon-Olivo/publication/327703478_Capitulo_1_Generalidades_de_las_redes_neuronales_artificiales/links/5b9fe3c0299bf13e6038a1d8/Capitulo-1-Generalidades-de-las-redes-neuronales-artificiales.pdf. pp 18-20.

[34] E. Acevedo, A. Serna, E. Serna, academia.edu, 2017. [Online]. Available: https://www.academia.edu/39630373/DESARROLLO_E_INNOVACI%C3%93N_EN_INGENIER%C3%8DA_Editorial_IAI. pp 175-180.

[35] S. Pattanayak, Springer, 2017. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4842-3096-1. pp 179-187. pp 1

[36] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn y D. Yu, IEEEXplore, 2014. doi: https://doi.org/10.1109/TASLP.2014.2339736

[37] C. Rincon, Universidad Politecnica de Madrid, 2007. [Online]. Available: https://lorien.die.upm.es/barra/pfcs/2007-carmenr/docs/proyecto.pdf.

[38] A. Nogueira, Universidad Federal do Amazonas, 2008. [Online]. Available: https://tede.ufam.edu.br/bitstream/tede/2959/1/DISSERTACAO%20ADRIANO%20NOGUEIRA.pdf.

[39] L. Valente, Universidad de Castilla - La Mancha, 2017. [Online]. Available: https://ruidera.uclm.es/xmlui/bitstream/handle/10578/15422/TFG_LUISALBERTOVALENTE.pdf?sequence=1.

[40] C. Luna, I. Bevacqua, N. Salvay, Universidad Tecnologica Nacional, 2011. [Online]. Available: https://www.profesores.frc.utn.edu.ar/electronica/fundamentosdeacusticayelectroacustica/pub/file/FAyE0711E1-Luna-Bevacqua-Salvay.pdf.

[41] D. Ginestar, Universitat Politecnica de Valencia, 2022. [Online]. Available: https://personales.upv.es/dginesta/docencia/posgrado/sparse.pdf. pp 15-21.

[42] V. Roman, Ciencia & Datos, 2019. [Online]. Available: https://medium.com/datos-y-ciencia/introduccion-al-machine-learning-una-gu%C3%ADa-desde-cero-b696a2ead359. pp 1

[43] R. Hernández, E. Pérez-Perdomo, D. Orozco, L. Sánchez, ResearchGate, 2018. doi: 10.13140/RG.2.2.26893.84961

[44] S. Uddin, A. Khan, E. Hossain, A. Moni, ResearchGate, 2019. doi: “https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8“ \t “_blank“ 10.1186/s12911-019-1004-8

[45] J. Martinez, IArtificial.net, 2020. [Online]. Available: https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/.