Automatic Phonetic Segmentation of the Yuhmu Language Using Mel Scale Spectral Parameters

Eric Ramos-Aguilar, J. Arturo Olvera-López, Ivan Olmos-Pineda, and Ricardo Ramos-Aguilar

***Abstract*—The application of digital signal processing techniques and machine learning, along with implicit segmentation, poses a challenge in the study of phonetic segmentation of indigenous languages in Mexico, given their linguistic and phonetic diversity. The analysis of Mel-scaled spectrograms offers an effective approach to identify patterns that can outline relevant information. By comparing the results with the actual number of phonemes in a word, both successes and areas for improvement can be observed. This article proposes a methodology for automatic segmental analysis of the Yuhmu language, considering parameter search in the Mel scale and implementing the cosine distance between spectrogram vectors. Additionally, relevant data within the resulting matrices are taken into account based on four key thresholds in information selection. The analysis yields a Segment Error Rate (SER) ranging from 38.79% to 41.35%, which aligns with the results reported in the literature on the subject.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9659

***Index Terms*—Implicit Segmentation, Phoneme Analysis, Low-Resource Language, SER, Yuhmu Language.**

I. Introduction

CURRENTLY, there exist more than 7,000 spoken languages in the world [1], organized into families that group linguistic components, making them unique across the globe. Among these, the Indo-European and Sino-Tibetan families dominate, which account for a significant number of speakers and are primarily studied by researchers. Languages such as Spanish, English, Mandarin, and Arabic are notable examples.

Today, there are various automatic language analysis approaches that offer ways to understand the language at a computational level. These approaches are integrated into Natural Language Processing (NLP), where the segmental and suprasegmental descriptors of phonetics (articulatory, phonetic, and auditory) are studied [2]. These descriptors are relevant for tonal and phoneme analysis.

Phonetic and tonal analysis has contributed to understanding and providing feedback on the correct pronunciation of words, sentences or phrases by nonnative speakers of a target language without relying on human evaluators to listen and analyze.

The use of computational tools has enabled the exclusive identification of linguistic elements for analysis. However, the models and methods mentioned require a substantial amount of data to be automatically trained. At present, many languages lack enough digital resources for proper analysis, which poses a significant challenge to researchers. These languages are considered Low-Resource Languages (LRLs) [3] and often include indigenous languages from all over the world.

Mexico is rich in cultural and linguistic diversity, particularly in indigenous languages, encompassing eleven language families such as: Álgica, Yuto-nahua, Cochimí-yumana, Seri, Oto-mangue, Maya, Totonaco-tepehua, Tarasca, Mixezoque, Chontal from Oaxaca, and Huave. According to the National Institute of Indigenous Languages, these families include 68 indigenous languages, with Nahuatl, Tseltal, Mixteco, Tsotsil, Zapoteco, Otomí, Totonaco, Chol, and Mazateco standing out due to their large number of speakers [4].

The preservation of Mexico's Indigenous Languages (MILs) is crucial for safeguarding the identity and culture of these communities, maintaining their original forms of communication, and ensuring the survival of their unique linguistic heritage. While computational advancements have been widely applied to Computer-Assisted Pronunciation Training (CAPT), creating tools that use automatic speech recognition for pronunciation assessment and learning evaluation, these technologies have not yet been extended to MILs [5]. A major challenge is the lack of linguistic databases (as training data) containing phonetic transcriptions and voice recordings. To improve accuracy in voice evaluation, voice segmentation is essential, breaking speech into its smallest units, phonemes, which is critical for tasks like voice synthesis, speech and language recognition, and speaker verification [6].

The use of phonetic segments is beneficial for identifying correct pronunciation, which can be accomplished in two ways: manually or automatically. In manual segmentation, human experts, typically linguists, temporally align the continuous digital audio waveforms to discrete phonetic symbols [6]. This analysis has been implemented to perform a subjective analysis of indigenous languages, as seen in [7], which explains the prosody of Mixtec or in [9], which compares the tonality of Yuhmu. Additionally, [10] and [8] conduct analyses of sound production in Nahuatl. However, this speech analysis incurs significant human and computational costs.

On the other hand, automatic segmentation is categorized as: explicit or implicit. Explicit segmentation involves textual elements specifically aligned within the digital audio, where

The associate editor coordinating the review of this manuscript and approving it for publication was Carlos Thomaz (*Corresponding author: Eric Ramos-Aguilar*).

Eric Ramos-Aguilar, J. A. Olvera-López and I. Olmos-Pineda are with Autonomous University of Puebla, Puebla, Mexico (e-mails: eric.ramosaguilar@viep.buap.mx, jose.olvera@correo.buap.mx, and ivan.olmos@correo.buap.mx).

R. Ramos-Aguilar is with the National Polytechnic Institute, Tlaxcala, Mexico (e-mail: rramosa@ipn.mx).

a model learns features of the speech segment using forced aligners. However, this type of segmentation is language-specific then it is not enough general when applied to other languages, particularly speech recognizers are typically not optimized for speech segmentation tasks [6]. This can be considered supervised segmentation due to the extensive number of labels available for performing segmentation.

Unsupervised segmentation involves fewer elements incorporated within the digital audio, and it is based on the premise that the voice characteristics within the phonetic segment exhibit stability. This type of segmentation is referred to as implicit, blind or text-independent. It identifies patterns through distributions or changes in properties to infer where divisions should be made. Unlike explicit segmentation, data labeling or text alignment for training is not required, which allows it to work with different languages that lack transcription or parallel translation information.

Currently, to our knowledge, there are not records of explicit and implicit phonetic segmentation analysis in Indigenous Languages of Mexico (LIM). Existing works have been manually developed using freely accessible tools like PRAAT or ELAN for study [7], [8].

To perform implicit segmentation, this work proposes the contrast identification in phonetic information within Mel-scale spectrograms via cosine distance. Additionally, a threshold is considered to find relevant points that assist in locating segments between each phoneme of a word. Therefore, Section 2 describes some related works on the topic. Section 3 details the database used in the research. Section 4 explains the methodology employed, outlining the steps taken to obtain the Mel-scale spectrograms and the process leading to segmentation. The results obtained are presented in Section 5. Finally, Section 6 includes the conclusions and future directions from the proposed work.

II. Related Work

The development of automatic segmentation has advanced across various fields, identifying effective methods and features for analysis. Language analysis using neural networks requires processing large datasets for training and validation, primarily focusing on digital audio, with transcriptions often used for support. This involves aligning audio and text, with speech segmentation referring to the alignment of phonemes, the smallest sound units [11]. In [12], a model for acoustic and phonetic decoding (APDM) automatically recognized pure and emphatic vowels in Modern Standard Arabic using Genetic Algorithms (GAs) for optimization efficiency and reduced computational cost. Techniques such as Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coding (LPC) were used to extract vocal tract coefficients, and the Manhattan distance was applied to classify vowel segments. With a corpus of 46,000 data points from Algerian speakers in noisy environments, the model achieved a 98.02% average classification rate.

In [13], automatic phonetic segmentation addressed intra-phoneme variations and inter-phoneme similarities in continuous speech using the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) framework. The main segmentation errors were from diphthongs and the boundaries between glides and vowels. To address these, two approaches were used: state-level model selection for dissimilarity within the same phoneme and context-dependent models for co-articulation. The topology of the HMMs was also adjusted considering phoneme duration. The segmentation accuracy was 91.32% in 20 ms on the TIMIT corpus, with a relative error reduction of 3.34%.

In [14], a performance trade-off is identified between phoneme categorization, phoneme segmentation, and word segmentation across various self-supervised learning algorithms based on Contrastive Predictive Coding (CPC). The experiments showed that networks that build context, while necessary for good categorization, hinder segmentation by causing a temporal shift in the learned representations. To address this issue, the authors proposed mACPC (multi-level Aligned CPC), a variant of ACPC that incorporates modeling and optimization at multiple levels to detect spectral changes. This approach improved performance across all tested categorization metrics and achieved state-of-the-art performance in word segmentation, using the LibriSpeech train-clean-100 dataset.

The enhancement of phoneme recognition in Arabic speech through the optimal selection of distinctive phonetic features utilizes a genetic algorithm to reduce the dimensionality of the acoustic feature vector. Since using extensive feature vectors increases computational complexity and affects real-time applications, a base model was constructed using feedforward neural networks to evaluate the performance of the proposed method. Experiments conducted with the Arabic Phonetic Database from King Abdulaziz City for Science and Technology demonstrated that the genetic algorithm-based method achieved slightly superior recognition accuracy compared to the full vector method, reaching a 50% reduction in input vector dimensionality and a recognition accuracy of 90%. Additionally, the results were validated using the Wilcoxon signed-rank test [15].

In [16], a neural architecture with a parameterized structured loss function was proposed for phoneme boundary detection, achieving state-of-the-art performance on the TIMIT and Buckeye corpora without phoneme input, outperforming baseline models in F1 and R metrics. Introducing phonetic transcription as additional supervision led to slight performance improvements and better convergence rates. The model was also tested on a Hebrew corpus, showing that phonetic supervision benefits multilingual contexts. In [17], a text-independent phonetic segmentation method using speech features from wavelet packet decomposition and a Sparse Representation Classifier (SRC) was developed, outperforming Mel-Frequency Cepstral Coefficients (MFCC) in speech segmentation tasks across English (TIMIT) and Arabic datasets. The SRC classifier also showed a higher hit rate than the k-Nearest Neighbors (k-NN) classifier, suggesting the proposed method's effectiveness in phonetic segmentation.

In [19], the problem of phoneme segmentation involves identifying the boundaries of phonetic units, primarily in English and other languages with abundant datasets. Popular

techniques, such as Hidden Markov Models and deep neural networks, require large volumes of data for training, which poses challenges for low-resource languages. In response, the work proposes a single-scan segmentation method based on geometric quadrilaterals, enabling the identification of boundary points by locally inspecting the shape of the speech signal. This method is efficient as it identifies boundaries in a single inspection, utilizing the geometric nature of waveform trajectories and treating the input signal as a sequence of structural components. The segmentation process was evaluated with spoken numbers in Indian-accented English and sentences in Telugu, demonstrating the effectiveness of the proposed approach.

The study proposed in [6] generates a strategy for automatic speech segmentation based on cosine distance similarity to identify phoneme boundaries. This strategy facilitates the selection of appropriate feature extraction techniques for speech segmentation. The authors developed a new combination of vocal tract features, termed FICV (Forward and Inverse Characteristics of Vocal Tract). The results showed that the proposed technique achieved a 14.48% error rate and an 85.2% accuracy with an alignment error within 10 ms, surpassing current techniques by 12.29% in error rates and 22.73% in alignment accuracy, highlighting the potential of FICV in speech segmentation.

The use of machine learning for phonetic segmentation generation has been recurrent in the aforementioned studies; on the other hand, the use of large amounts of digital audio for its implementation is a limitation considered by the LIMs. However, it is possible to automate the segmentation and supervised learning processes. This paper analyzes the implicit segmentation of phonemes in the Yuhmu language, which does not consider a textual transcription; representing a subsequent practice to that presented in [20], where Praat is used and the segmentation is done manually. The aim of this process is to achieve automatic phoneme segmentation aided by signal processing techniques.

III. Yuhmu Language

The Yuhmu language is a variant of Otomi (a macrolanguage of the Otomangue linguistic group spoken by an ethnic and cultural group distributed in the central-southern region of Mexico), located in the municipality of Ixtenco, Tlaxcala, Mexico. This language is endangered, as only a few elderly speakers (approximately 70 years old) maintain its pronunciation. Some individuals under 60 years old understand the language, but there are no children learning Yuhmu as their native language [20], leading to a decline in the language with the prospect of its extinction.

According to a community census conducted by [20], there are around 100 speakers of Yuhmu, although their level of linguistic competence is not well-documented. The language lacks a native writing system, which has led to efforts to phonetically represent its sounds by developing isolated writing systems incorporated by various historians.

Yuhmu consists of 32 phonemes classified according to the International Phonetic Alphabet (IPA), including 12 vowels (V) that can be either oral or nasal, as illustrated in Fig. 1. It also includes 20 consonants (C) categorized based on the location of the articulatory organs in the vocal tract. Unlike vowels, consonants can be voiced or voiceless.

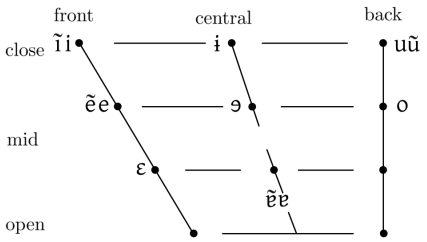


Fig. 1. Phonetic symbols of vowels (those with a tilde above them are considered nasal, while those without it are oral).

In Table I, as a summary, the phonetic representation of the consonants of Yuhmu are shown. This table contains two general columns, each describing more specific features:

- *Airway obstruction mode*: This describes the way air flows through the vocal tract, naming this passage with terms such as plosive, affricate, fricative, nasal, among others, and whether they produce vibration within the glottis (voiceless or voiced).
- *Airway obstruction site*: The specific location or area in the respiratory tract where an obstruction occurs that impedes the normal airflow from the lungs. According to the position of the tongue, lips or glottis, the obstruction can be named as bilabial, alveolar, palatal, velar or glottal.

TABLE I
Symbols of the International Phonetic Alphabet for Consonants in Yuhmu

Airway obstruction mode		Airway obstruction site: Bilabial	Alveolar	Palatal	Velar		Glottal
Plosive	Voiceless	p	t		k	k^w	ʔ
	Voiced	b	d		g	g^w	
Affricates	Voiceless		ts	tʃ			
Fricative	Voiceless		s	ʃ			h
	Voiced		z				
Nasal	Voiced	m	n				
Tap or Flap	Voiced		ɾ				
Approximant	Voiced			j	w		

For the analysis, 330 digital audio recordings of word pronunciations were considered, including three repetitions per word. These recordings encompass phonetic combinations characteristic of Yuhmu, organized according to the patterns C-V, C-C-V, C-C-C-V, and C-V-V-V. These combinations can appear at the beginning, middle or end of words, and tonal variations such as high, low, and low-high tones are also taken into account [20].

Recordings were conducted using Audacity, an open-source software widely used in fieldwork, with a BEHRINGER C1-U condenser microphone (frequency response from 40 Hz to 20 kHz and a maximum sound pressure level of 136 dB) that ensures clear and accurate capture of the vocal signal. Audio files were stored in monophonic WAV format, an uncompressed standard that preserves the original sound quality crucial for subsequent acoustic and computational analyses [21], [22].

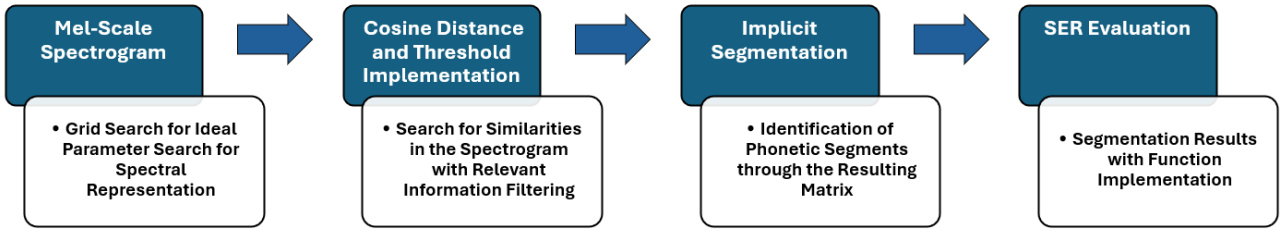


Fig. 2. Methodology phases for phonetic segmentation of Yuhmu.

The recorded words were selected from a base dictionary organized into semantic fields such as family, fruits, food, household and field elements, animals, and body parts, thus ensuring representative coverage of the everyday Yuhmu lexicon and a balanced distribution of phonemes. Recording sessions took place in the speakers' homes, aiming for controlled environments to minimize background noise and to obtain high-quality acoustic samples.

The base dictionary used for words in Yuhmu is the one proposed by [20], which describes all the phonemes incorporated in the language and is analyzed subjectively. Digital recordings vary in duration from 376 ms to 1.118 s and undergo preprocessing that includes noise attenuation, amplification, and word trimming. These recordings include a transcription of the phonemes presented in each word, which do not follow a conventional writing system but instead represent the written form of the pronunciation. The dataset used in this study is not publicly available and cannot be shared due to ownership restrictions. However, it may be available upon reasonable request, under specific conditions.

IV. Methods

In this section, the proposed method for generating implicit phonetic segmentation (IPS) of the Yuhmu language is described. This method is based on the analyzed literature, from which relevant information is extracted using a Mel-scaled spectrogram. Subsequently, operations are performed on the resulting spectrogram matrix to apply a threshold, allowing for further processing of the resultant matrix to automatically identify segments. The proposed method consists of four phases (see Fig. 2), which are detailed in the following sections.

A. Mel-Scale Spectrograms

The spectral representation of digital audio in Mel-scale spectrograms considers relevant analytical aspects, recognizing that variations in the parameters contribute to improving the efficiency and effectiveness of phonetic segmentation. It is important to note that by modifying the window size, the number of triangular filters (Mels) or spectral overlap, the results vary with different combinations. Fig. 3 illustrates a diagram of the process that digital audio must undergo to obtain a Mel-scale spectrogram.

Therefore, a comprehensive search for parameters is conducted using the Grid Search method, which is a hyperparameter optimization technique in machine learning models.

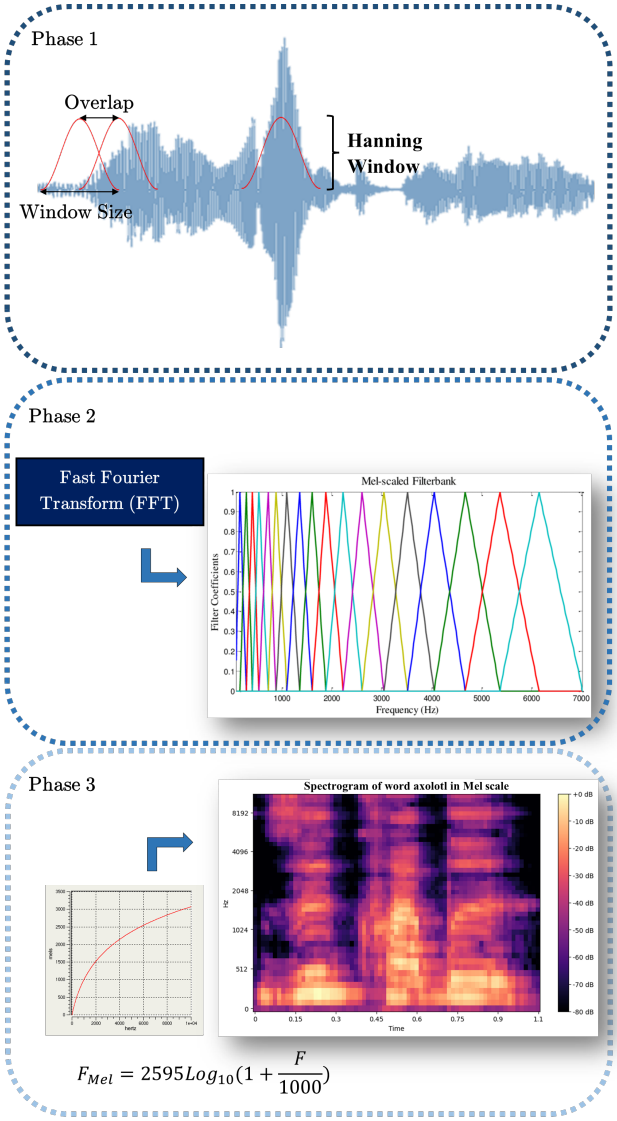


Fig. 3. Process of converting digital audio to a Mel scale spectrogram.

This technique involves defining a set of values for each hyperparameter and evaluating all possible combinations to find the one that offers the best performance [24].

The search for parameters to find the optimal spectral representation of the spectrogram is generated from the proposal shown in Table II. This technique, employed in signal processing, allows for the analysis of short segments of an audio signal. The audio signal is divided into short segments, known as windows, which overlap to capture both temporal

and frequency information. Each window undergoes a Fourier analysis to convert it from the time domain to the frequency domain, enabling the visualization of the signal's energy at different frequencies during that temporal segment [25]. The Mel

TABLE II
SELECTED PARAMETERS FOR OPTIMAL MEL-SCALE SPECTROGRAM SEARCH

Window	Window size (ms)	Overlap (%)	Mel filter bands
Hanning	20, 25, 30, 35, and, 40	25, 50, and, 75	15, 20, 25, 30, 35, 40, and, 45

scale is a perceptual frequency scale for humans, calculated through a nonlinear transformation from frequency in Hertz to the Mel scale. This can be performed using equation 1 (where F represents the frequency of the digital audio signal). For this study, reference is made to [26], which provides relevant information for analyzing the Yuhmu language, considering the Hanning window as optimal for this analysis.

$$F_{Mel} = 2595 Log_{10}(1 + \frac{F}{1000}) \tag{1}$$

In the literature, it is recommended that the window size should range from 20 to 40 ms, with increments of 5 ms to ensure an adequate interval for analysis [23]. The overlap is set at 25%, 50%, and 75% of this window size, which helps minimize information loss during each analysis.

The window size determines the temporal and frequency resolution of the analysis; a range of 20 to 40 ms is common in speech processing because it allows capturing dynamic features without losing temporal information or spectral resolution. Increments of 5 ms provide precise adjustment of the window according to the study's needs. Additionally, the choice of the Hanning window is appropriate because it attenuates side lobes in the frequency domain, reducing spectral leakage and improving the clarity of relevant acoustic components.

Regarding overlap, 50% is a typical value that balances temporal resolution and artifact reduction, but 25% or 75% can be applied depending on the signal characteristics and analysis goals to preserve information at the edges of the windows.

Although the literature recommends between 15 and 20 Mel bands (Mels, filters that simulate human auditory perception) [18], a wider range of 15 to 45 Mels is proposed to capture more information in the spectrogram and its resulting matrix. This increase is especially relevant for languages which have not been widely studied computationally, as it may reveal unique acoustic features.

Since Mels correspond to rows in the matrix, a higher number of Mels the better detailed representation, facilitating better discrimination of phonemes and subtle linguistic elements, which supports analysis and recognition tasks in underexplored languages.

B. Cosine Distance and Threshold Implementation

The Cosine Distance (Equation 2) is applied row-wise to the resulting spectrogram matrices. This analysis allows for the evaluation of similarity between the temporal evolutions (i.e., across time frames or columns) of each frequency component (Mel band) within the spectrogram. By comparing the spectrogram data element-wise along the rows, the method helps to identify consistent activation patterns over time and contributes to the grouping of information, thereby facilitating the segmentation associated with each phoneme.

It is important to highlight that the row-wise analysis involves pairing data column by column, as each operation evaluates pairs of elements corresponding to the same temporal positions (columns) in both rows being compared. This establishes a contrast of information (Fig. 4) based on the digital audio representation and the segmentation derived from it.

The cosine distance is defined as:

$$Cos\_D(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2}$$

Where:

- x_i and y_i are the values corresponding to the same frequency band (Mel band) across two different time frames (columns) of the spectrogram.
- n is the number of Mel bands (rows in the spectrogram matrix).
- The symbol $\cdot$ denotes element-wise multiplication between the vector components (row values).

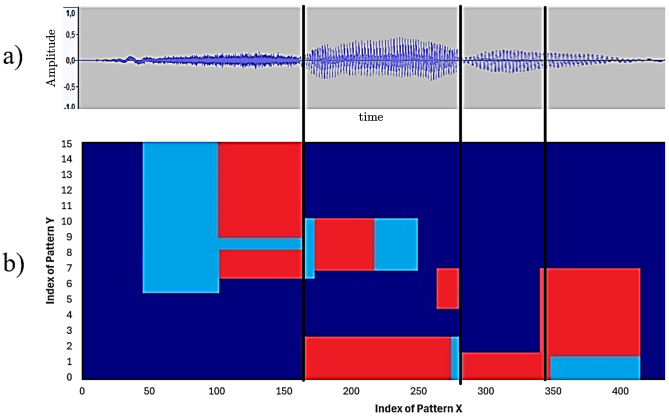


Fig. 4. a) Speech samples with phoneme boundaries over time. b) Hypothesis for ideal scores (Where navy blue tones represent null information, red tones represent high information, and light blue tones represent low information).

After computing the cosine distance between the columns of the spectral matrix, a search for relevant information is performed for each resulting matrix. Upon examining the data in the matrix, there is information close to zero that introduces noise in the segment search, creating a significant contrast between phonetic information and these data points. Therefore, an information elimination process is applied using four different thresholds: 0.1, 0.2, 0.3, and 0.4 (considering that the energy of the voice varies in each case, however, the representation of the phonemes in a word exceeds these thresholds, which helps to filter and observe the phoneme separately). This information is removed, and zeros are placed in all of these points across all resulting matrices, resulting in sets of information in the matrix that create a data separation.

This can be represented in an image to observe the variation in the information (see Fig. 5).

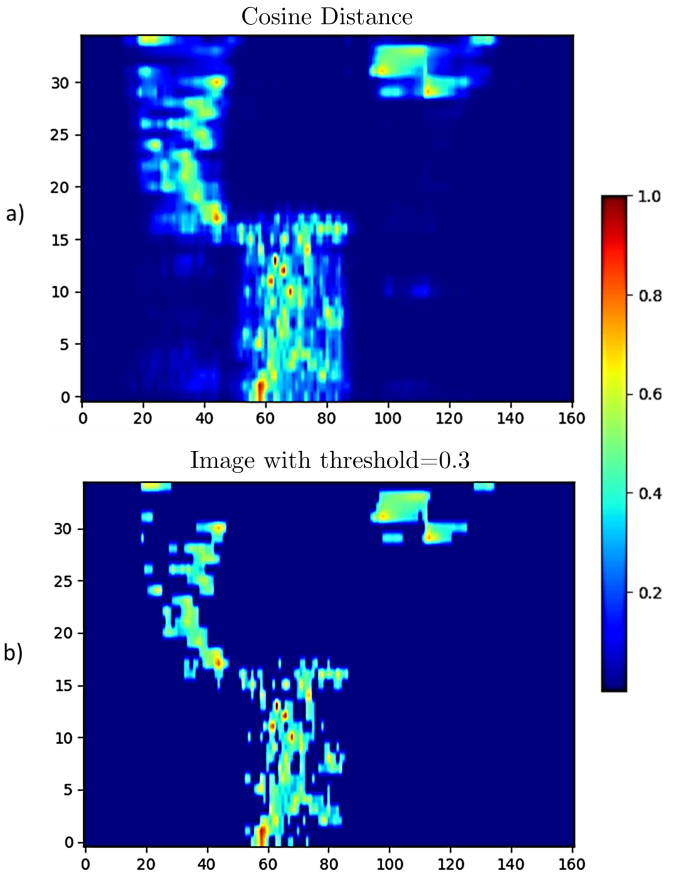


Fig. 5. a) Resulting matrix after applying cosine distance to the Mel-scale spectrogram. b) Threshold application on that matrix (Fig. 5a). The data are normalized: high values near 1 (red), low values near 0 (light blue), and zeros shown in navy blue.

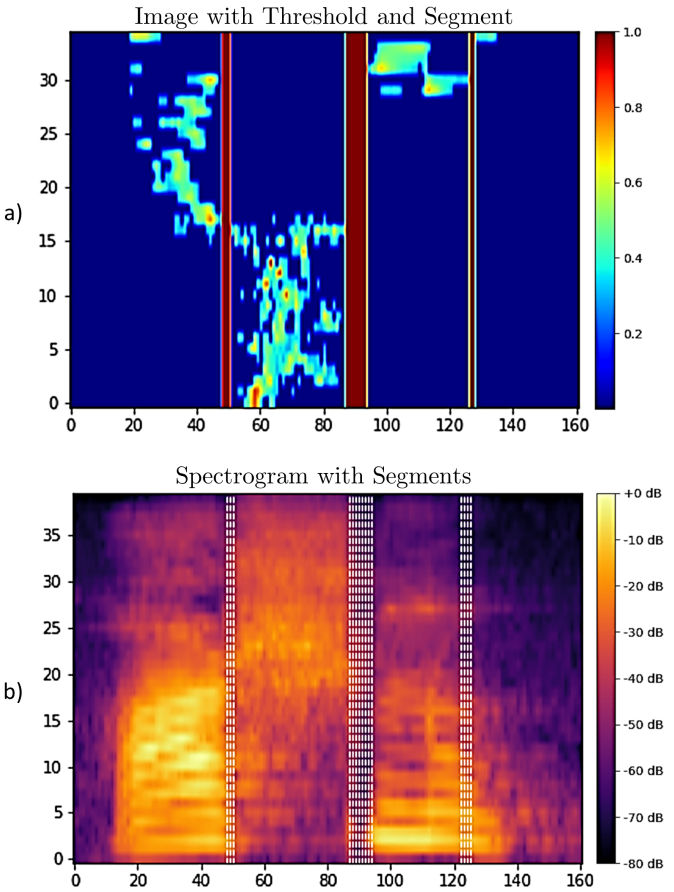


Fig. 6. a) Segments highlighting zero columns. b) Overlap of segments (dashed lines) on the thresholded Mel-scale spectrogram. The data represent the signal's spectral power on a logarithmic scale from 0 to -80dB, where 0dB indicates maximum power and -80dB indicates minimal spectral presence.

C. Implicit Segmentation

Taking into account the information given in the resulting matrix after applying the threshold, the zero columns that integrate in the middle of each automatically found data set are highlighted. This allows for a counting of these sets, resulting in a segment in each case. It is important to mention that the initial and final zero columns are not included, as they would skew the segment counting process within the resulting matrices. After this, an overlap of the found segments with the spectrogram is performed, which helps to delineate the spectral information and find the phonetic representation of each word (see Fig. 6).

The visual representation in phonetic segmentation is of significant importance, as the generation of IPS for the Yuhmu language can provide a visualization of each of the segmented phonetic representations. Taking this into account and making a comparison with the research considered in this paper, where only the resulting statistical aspects are shown, IPS allows for the consideration of a statistical outcome of the phonemes found per word, but it also provides a visual representation of the spectrogram's separation in an image. This helps in making a comparison between the image and the digital audio representation for subsequent sound segmentation.

D. Evaluation Using Segment Error Rate (SER)

In the implicit analysis, an automatic search for detected segments is performed. Within the database, the number of phonemes contained in each word is known; therefore, when evaluating the segmentation, it is possible to compare the number of segments found by the method with the actual number of segments in each word. This comparison is quantified using equation 3, which defines the Segment Error Rate (SER) as:

$$SER = \frac{\#error\_segments}{\#reference\_segments} \tag{3}$$

where $\#error\_segments$ is the absolute difference between the total number of reference segments and the number of detected segments; $\#reference\_segments$ is the total number of segments within each word.

SER is a fundamental metric for assessing the quality of phonetic segmentation, as it objectively quantifies the discrepancy between the number of segments detected by an automatic method and the true number of phonemes composing a word. This metric is particularly valuable as it

provides a clear and easily interpretable indicator of overall system performance, facilitating comparison across different configurations or algorithms.

SER is based on a straightforward formula that reflects the relative error in the number of segments, making it practical for evaluating large datasets and numerous parameter combinations without requiring complex manual intervention. Its simplicity enables rapid and efficient analysis, which is crucial in automated optimization processes where measuring the impact of hyperparameter variations in a scalable manner is necessary.

Although SER does not explicitly account for temporal precision or the exact localization of phonetic boundaries, its strength lies in offering a robust initial evaluation of segmentation errors. It serves as a solid foundation for parameter selection and system improvement. Subsequently, this metric can be complemented using methods that assess temporal accuracy to achieve a more detailed and comprehensive evaluation.

V. Results

An exhaustive analysis of the parameters generated a total of 95,040 results, yielding an average segment error rate (SER) of 54.94% across all possible parameter combinations. Following the general analysis, a refined selection of specific parameters was made to improve the SER outcome. This selection takes into account the number of filters, the threshold applied to the resulting cosine distance matrix, and the overlap used in obtaining the Mel spectrogram.

Table III presents the findings after selecting relevant parameter combinations, achieving an improvement in SER of up to 16.741%. Specifically, by setting 40 filters, a threshold of 0.4 and an overlap of 75%, with a window size of 20 ms a SER of 38.799%.

The high SER value is explained by an excessively detailed spectral segmentation, that is, a division of the spectrum greater than expected considering the number of phonemes present, caused by threshold values that, when combined, become dynamic in each set of parameters. This leads to a significant loss of information, reflected in a higher number of columns with values close to zero in the spectral representation. As a result, the number of detected segments increases, which leads to a higher error due to the larger number of segments that must be predicted.

In addition to improving the SER, it can be observed that the best results occur when a threshold of 0.4 and 75% overlap are implemented, regardless of the filter combination.

The dynamic nature of the hyperparameter values allows for the identification of combinations that optimize the phonetic segmentation of recorded words, generating multiple spectral representations per word (see Table IV). In total, 18,271 results with a 0% Segment Error Rate (SER) were obtained, reinforcing the validity of the approach since each base word, with three repetitions, can have more than one valid segmentation.

These results include the entire base dictionary, consisting of 330 words with three repetitions each. The hyperparameter search provides at least 55 segmented spectrograms per word with different configurations, demonstrating that the 0% SER is not dependent on a single combination. Moreover, by integrating all words, the full phoneme inventory of the dictionary is covered, with expanded spectral representation since segmentation occurs across multiple spectra per word.

TABLE III
Results Obtained from the Set of Analyzed Hyperparameters and the Average Results in SER (%)

Number of filters and threshold	Overlap %		
	25	50	75
10,0.1	46.0866	43.8	47.5559
10,0.2	48.27	51.11	54.3663
10,0.3	53.6769	44.7681	43.12
10,0.4	65.1018	46.9389	**41.2194**
15,0.1	50.5612	51.6858	56.3749
15,0.2	47.0607	44.9615	49.814
15,0.3	53.25	42.84	44.05
15,0.4	82.2451	48.8796	**39.8739**
20,0.1	52.1296	53.8207	57.2906
20,0.2	47.3727	44.9396	50.13
20,0.3	53.6497	42.2494	45.3975
20,0.4	93.8297	52.3784	**39.4621**
25,0.1	53.7815	54.1095	58.3827
25,0.2	47.9824	48.373	51.127
25,0.3	55.0396	41.8386	43.3201
25,0.4	101.5236	50.7013	**39.1117**
30,0.1	50.5612	51.6858	60.1452
30,0.2	47.3727	44.9396	51.87
30,0.3	57.0396	43.8386	44.4828
30,0.4	111.5236	53.7013	**39.4449**
35,0.1	54.19	55.78	60.69
35,0.2	50.46	47.64	51.83
35,0.3	63.76	43.7249	56.46
35,0.4	123.76	56.4617	**39.9797**
40,0.1	57.51	56.67	61
40,0.2	49.4337	47.94	52.31
40,0.3	63.08	42.5791	45.06
40,0.4	125.0311	55.27	**38.799**
45,0.1	63.32	58.85	60.59
45,0.2	56.554	50.94	50.86
45,0.3	72.89	45.55	43.34
45,0.4	141.79	59.14	**41.35**

The dynamic nature of the hyperparameter values allows for the identification of different combinations that yield the best results. This enables the correct phonetic segmentation of a recorded word, generating distinct hyperparameter representations for each word. It is important to note that this 0% SER was found in 18,271 segmented results with zero SER, making the representativeness significant, since the dynamic values allow for obtaining segmented spectral representativeness; furthermore, each base word has three repetitions, and each recorded word can have more than one segmented spectral representation.

When comparing with some of the state-of-the-art works presented, it can be observed that the accuracy of phonetic segmentation is above 90%. However, this segmentation is supported by text (explicit segmentation), as in the cases

TABLE IV
EXAMPLE OF DYNAMIC HYPERPARAMETERS WITH 0% SER FOR SOME WORDS SPOKEN IN YUHMU

Audio name in English	Number of filters	Overlap %	Threshold	SER %
Bee	45	75	0.3	0
Bee	45	75	0.4	0
Cane	10	25	0.2	0
Cane	15	50	0.3	0
Cloud	40	75	0.3	0
Cloud	45	50	0.4	0
Path	10	25	0.4	0
Path	15	25	0.4	0
Potato	10	50	0.4	0
Potato	15	25	0.3	0
Sing	10	25	0.1	0
Sing	15	75	0.4	0

of [12]–[15]. On the other hand, compared to [6], which considers an experiment with implicit segmentation using FICV, a phoneme error rate of approximately 22% is reported. This study focuses on the Arabic language, which has been consistently analyzed. Additionally, in [27], phoneme recognition is performed through audio for low-resource African languages. When averaging these results, a phoneme error rate (PER) between 25-78.4% is obtained. However, when African languages are analyzed separately, the model did not predict direct phonemes, using 2,321k recordings for training.

Based on the information above, it can be stated that the results obtained are significant compared to the studies mentioned. The relevance becomes even more pronounced when considering that the computational techniques applied in our research are completely independent of machine learning, using basic computational resources for analysis, which reduces execution time and computational requirements.

The obtained results gain greater relevance by presenting phonetic segmentation through a visual analysis based on statistics derived from spectral representations. The integration of pattern recognition within the spectrogram enables a visual statistical analysis that facilitates a deeper and more detailed understanding of the phonetic segments, as well as the acoustic and phonological characteristics involved.

Furthermore, this approach aligns with a subjective evaluation performed by a linguist; however, the human cost is considerable, since the analysis was carried out by recording individually, unlike our method which is applied automatically. The subjective segmentation by the linguist matches the obtained results, although there is no representation of the temporal variation of the phoneme in milliseconds within the segments. Nevertheless, this tool becomes fundamental for analysis and to verify the segmentation, since even though only the phonemes present are known, there is no detailed record of the exact boundaries between them, thus helping to generate a segmentation that represents an important advance in the analysis of the language.

To complement this visual approach, a quantitative statistical analysis was conducted to evaluate the impact of different parameters in spectral segmentation on performance, measured by the SER. Two complementary methods were applied: the non-parametric Friedman test and a repeated measures analysis of variance (ANOVA), considering three variables: overlap, threshold, and number of Mel filter bands. The results of these analyses are summarized in Table V.

TABLE V
STATISTICAL RESULTS FOR PARAMETERS EVALUATED ON SER

Parameter	Comparison	Method	Statistic	p-value	Pairs
Overlap	110 vs 220	Friedman	0.70	0.7047	37
	vs 330	ANOVA	2.74	0.0689	–
Threshold	0.1 vs 0.2	Friedman	8.02	0.0456	29
	vs 0.3 vs 0.4	ANOVA	5.03	0.0026	–
Mel Filter Bands	10 vs 20	Friedman	5.00	0.1718	12
	vs 30 vs 40	ANOVA	1.68	0.1811	–

The threshold parameter demonstrated a statistically significant effect on SER in both the Friedman and ANOVA analyses. This result is consistent with the nature of the thresholding process applied to eliminate low-energy information from spectral matrices. By using threshold levels of 0.1, 0.2, 0.3, and 0.4, regions with low energy often corresponding to silences or background noise are removed, while regions with higher energy, typically associated with phonemes, are preserved. This operation inserts zeros at the positions below the threshold, effectively segmenting the Mel spectrogram and enhancing phoneme distinction, which leads to a lower SER.

In contrast, overlap and the number of Mel filter bands did not produce statistically significant differences in SER across the tested ranges. Despite the lack of significance, these parameters still play a fundamental role in the structure of the spectral representation: overlap influences temporal resolution by controlling the window advance during framing, and the number of Mel filter bands determines the granularity of the frequency resolution. The results suggest that, within the evaluated settings, their variation does not critically affect SER, although they remain essential for the signal representation process.

As part of a supporting analysis, an experiment was conducted using a method based on acoustic embedding representations to complement previous studies.

This method employs a simple convolutional encoder that processes Mel spectrograms from audio files in wav format, generating embedding vectors for each temporal segment. By computing the cosine similarity between consecutive embeddings, abrupt drops can be identified and interpreted as potential phonemic boundaries. This unsupervised implicit segmentation enables the automatic detection of phonetic boundaries without requiring manual annotations [28].

TABLE VI
EXAMPLES OF PHONEMIC BOUNDARIES DETECTED USING ACOUSTIC EMBEDDING REPRESENTATIONS

Word	Boundaries (seconds)	Number of Segments
bee	—	0
grandmother	0.05; 0.08; 0.11	3
inside	0.22; 0.24; 0.27; 0.29; 0.31	5
water	0.02; 0.05; 0.09; 0.50	4
avocado	0.04; 0.09; 0.13; 0.17; 0.22; 0.26	6

In addition, a clustering analysis using K-means was applied to the resulting Mel-scale spectrograms. This analysis groups acoustically similar segments; however, it is specifically based on the power displayed in the spectrogram matrix. Depending on the energy distribution, clusters are formed. Nevertheless, the clustering does not segment the required information into phonetic spaces, and therefore, it does not perform temporal segmentation within the spectrogram. Its use is complementary rather than a substitute. The values of k vary between 2 and 8, depending on the spectrogram, as shown in Fig. 7.

a) b) c)

Fig. 7. Representation of segment search using K-means, with various words spoken in Yuhmu, a) k = 5, b) k = 7, and c) k = 8.

The results of this experiment, which report boundary times in seconds and the number of segments per file (see Table VI), do not constitute a comparative analysis with our main methodology. Our approach integrates experimental statistical and visual analysis and is specifically designed for Indigenous languages such as Yuhmu, which exhibit unique acoustic properties not considered in existing machine learning methods developed for non-tonal languages.

Therefore, this experiment serves as a complementary and standalone analysis for phonetic studies in Indigenous languages of Mexico.

VI. Conclusions

Segmental phoneme analysis is crucial for researchers, as it involves both phonetic and articulatory aspects. In the case of Yuhmu, a tonal language spoken in Ixtenco, Tlaxcala, its study presents significant challenges due to its unique variant, the lack of a standardized writing system, and its classification as a low-resource language in computational environments.

Optimal parameters yielding a zero Segment Error Rate (SER) have been identified, allowing for the dynamic adjustment of hyperparameters to more accurately detect word segments.

Having a dedicated database, even without a formal writing system, enables the generation of pronunciation feedback. This opens the door to developing educational tools similar to Duolingo or Rosetta Stone, through spectral and inverse spectral analysis to identify phonemes in digital audio recordings.

Additionally, the use of acoustic embeddings allows for a more precise and discriminative representation of phonetic segments. This strategy is intended to be complemented with a phonetic spectral image database, which will associate visual representations with acoustic features. This will support the training of models for automatic segment detection and pronunciation evaluation. While perfect segmentation is not guaranteed, this approach facilitates the analysis of relevant patterns and supports the creation of a structured database for the preservation and teaching of Indigenous languages.

This approach not only advances the study of Yuhmu but also it can be extended to other Indigenous languages such as Jñatrjo (a variant of Mazahua), Tutunaku, and Nahuatl, which already have recordings available. In doing so, it strengthens efforts to conserve and promote the linguistic and cultural heritage of Indigenous communities.

References

[1] Europa Press, Los idiomas, en cifras: ¿cuántas lenguas hay en el mundo?, s.f., https://www.europapress.es/sociedad/noticia-idiomas-cifras-cuantas-lenguas-hay-mundo-20190221115202.html, Accedido: 01-octubre-2024.

[2] F. Lasi, A study on the ability of supra-segmental and segmental aspects in English pronunciation, Ethical Lingua: Journal of Language Teaching and Literature, vol. 7, no. 2, pp. 426–437, 2020. DOI:10.30605/25409190.222

[3] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2021, pp. 2545–2568, Jun. 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.201/. DOI: 10.18653/v1/2021.naacl-main.201

[4] Secretaría de Cultura, México es uno de los países con mayor diversidad lingüística en el mundo, s.f., https://www.gob.mx/cultura/prensa/mexico-es-uno-de-los-paises-con-mayor-diversidad-linguistica-en-el-\mundo#:~:text=Las%20lenguas%20que%20m%C3%A1s%20se,m%C3%A1s%20de%20200%20mil%20hablantes., Accedido: 01-octubre-2024.

[5] P. M. Rogerson-Revell, Computer-assisted pronunciation training (CAPT): Current issues and future directions, Relc Journal, vol. 52, no. 1, pp. 189–205, 2021. DOI:10.1177/0033688220977406

[6] M. Javed, M. M. A. Baig, and S. A. Qazi, Unsupervised phonetic segmentation of classical Arabic speech using forward and inverse characteristics of the vocal tract, Arabian Journal for Science and Engineering, vol. 45, pp. 1581–1597, 2020. DOI:10.1007/s13369-019-04065-5

[7] K. Penner, Prosodic structure in Ixtayutla Mixtec: Evidence for the foot, 2019. DOI:10.13140/RG.2.2.28786.96965

[8] R. Turnbull, The phonetics and phonology of lexical prosody in San Jerónimo Acazulco Otomí, Journal of the International Phonetic Association, vol. 47, no. 3, pp. 251–282, 2017.DOI:10.1017/S0025100316000384

[9] E. P. Velásquez Upegui, Entonación del español en contacto con el otomí de San Ildefonso Tultepec: Enunciados declarativos e interrogativos absolutos, Anuario de Letras. Lingüística y Filología, vol. 8, no. 2, pp. 143–168, 2020. DOI: https://doi.org/10.19130/iifl.adel.2020.24875

[10] A. S. Sahagún, Spanish VOT Production by L1 Nahuat Speakers, PhD thesis, University of Saskatchewan, 2021.

[11] S. Brognaux and T. Drugman, Hmm-based speech segmentation: Improvements of fully automatic approaches, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 5–15, 2015. DOI: 10.1109/TASLP.2015.2456421

[12] M. Aissiou, A genetic model for acoustic and phonetic decoding of standard Arabic vowels in continuous speech, International Journal of Speech Technology, vol. 23, no. 2, pp. 425–434, 2020. DOI:10.1007/s10772-020-09694-y

[13] W. Peng, Y. Gao, B. Lin, and J. Zhang, A practical way to improve automatic phonetic segmentation performance, in 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 1–5, 2021, IEEE. DOI:10.1109/ISCSLP49672.2021.9362107

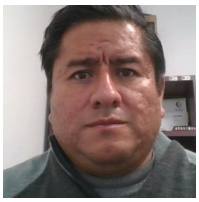
[14] S. Cuervo, M. Grabias, J. Chorowski, G. Ciesielski, A. Łańcucki, P. Rychlikowski, and R. Marxer, Contrastive prediction strategies for unsupervised segmentation and categorization of phonemes and words, in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3189–3193, 2022, IEEE. DOI: 10.1109/ICASSP43922.2022.9746102

[15] A. B. Ibrahim, Y. M. Seddiq, A. H. Meftah, M. Alghamdi, S. A. Selouani, M. A. Qamhan, Y. A. Alotaibi, and S. A. Alshebeili, Optimizing Arabic speech distinctive phonetic features and phoneme recognition using genetic algorithm, IEEE Access, vol. 8, pp. 200395–200411, 2020. DOI:10.1109/ACCESS.2020.3034762

[16] F. Kreuk, Y. Sheena, J. Keshet, and Y. Adi, Phoneme boundary detection using learnable segmental features, in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8089–8093, 2020. DOI: https://doi.org/10.48550/arXiv.2002.04992
[17] I. Al-Hassani, O. Al-Dakkak, and A. Assami, Phonetic segmentation using a wavelet-based speech cepstral features and sparse representation classifier, Journal of Telecommunications and Information Technology, no. 4, pp. 12–22, 2021. DOI: https://doi.org/10.26636/jtit.2021.153321
[18] Z. Kh. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022. DOI: https://doi.org/10.1109/ACCESS.2022.3223444
[19] P. Bhagath and P. K. Das, Quadrilaterals based phoneme segmentation technique for low resource spoken languages, in TENCON 2022-2022 IEEE Region 10 Conference (TENCON), pp. 1–6, 2022. DOI:10.1109/TENCON55691.2022.9977455
[20] R. Alarcon Montero, Manual para la escritura de los sonidos del Yuhmu, INAH, 2023.
[21] M. V. Belodedov, R. V. Fonkants, and R. R. Safin, "Development of an algorithm for optimal encoding of WAV files using genetic algorithms," in 2023 5th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE), vol. 5, pp. 1–6, 2023. DOI: https://doi.org/10.1109/REEPE57272.2023.10086837
[22] R. Liu, J. Zhang, and G. Gao, "Multi-space channel representation learning for mono-to-binaural conversion based audio deepfake detection," Information Fusion, vol. 105, p. 102257, 2024. DOI: https://doi.org/10.1016/j.inffus.2024.102257
[23] H. M. Fayek, "Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between," 2016. [Online]. Available: https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html
[24] J. Kim, Iterated grid search algorithm on unimodal criteria, Virginia Polytechnic Institute and State University, 1997.
[25] S. Pangaonkar and A. Panat, A review of various techniques related to feature extraction and classification for speech signal analysis, in ICDSMLA 2019: Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications, pp. 534–549, Springer Singapore, Singapore, 2020. DOI:10.1007/978-981-15-1420-3_57
[26] E. Ramos-Aguilar, J. A. Olvera-López, and I. Olmos-Pineda, A general overview of language pronunciation analysis based on machine learning, Research in Computing Science, vol. 152(10), 2023.
[27] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and others, Universal phone recognition with a multilingual allophone system, in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8249–8253, 2020. DOI: https://doi.org/10.48550/arXiv.2002.11800
[28] M. Buisson, B. McFee, S. Essid, and H. C. Crayencour, Self-supervised learning of multi-level audio representations for music segmentation, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 2141–2152, 2024. DOI: https://doi.org/10.1109/taslp.2024.3379894

Eric Ramos-Aguilar currently a Ph.D. candidate in Language and Knowledge Engineering in the Faculty of Computer Science, Autonomous University of Puebla. His primary research interests include natural language processing, pattern recognition, digital audio processing and analysis, and the study of indigenous languages of Mexico.

J. Arturo Olvera-López completed a PhD degree in Computer Science at National Institute of Astrophysic, Optic and Electronic (INAOE, Mexico). He is interested in problems related to the areas of pattern recognition, data mining, machine learning, data pre-processing, data reduction, digital image/signal processing & analysis, and biometrics.



Ivan Olmos-Pineda completed a PhD degree in Computer Science at National Institute of Astrophysic, Optic and Electronic (INAOE, Mexico). He is interested in problems related to the areas of pattern recognition, data mining, machine learning, data pre-processing, data reduction, digital image/signal processing & analysis, and biometrics.

Ricardo Ramos-Aguilar completed a PhD degree in Language and Knowledge Engineering in the Faculty of Computer Science, Autonomous University of Puebla. His primary research interests include natural language processing, pattern recognition, computer vision.