ISSN 1870-4069

Phonetic segmentation of the yuhmu language Using Mel-scale Spectral Representations

Eric Ramos-Aguilar[1,2], J. Arturo Olvera-López[1], Ivan Olmos-Pineda[1],
Barbara Emma Sánchez-Rinza[3], Ricardo Ramos-Aguilar[2]

[1] Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla,
Mexico

[2] Instituto Politécnico Nacional, UPIIT,
Mexico

[3] Facultad de Ciencias Físico Matemáticas,
Benemérita Universidad Autónoma de Puebla,
Mexico

eric.ramosag@alumno.buap.mx, {jose.olvera, ivan.olmos,
barbara.sanchez}@correo.buap.mx, rramosa@ipn.mx

Abstract. The study of phonetic segmentation in indigenous languages of Mexico poses a challenge due to their linguistic and phonetic diversity. The use of digital signal processing techniques, machine learning, and both implicit and explicit segmentation, along with Mel-scale spectrogram analysis, provides an effective approach to identifying patterns that may delineate relevant information. Comparing the results with the actual number of phonemes in a word reveals both successes and areas for improvement. This paper proposes a methodology for segmental analysis in language Yuhmu considering parameter search in Mel scale, implementing cosine distance between spectrogram vectors, taking into account relevant data within the resulting matrices and seeking patterns of interest. The segment error rate yields results ranging from 23.89% to 26.03%, close to those reported in the literature on the subject.

Keywords: Phonetic segmentation, audio analysis, indigenous language.

1 Introduction

The analysis of language for automatic speech recognition, speaker verification, or language identification takes into account various forms of study, but always employing a segmentation of phrases or words, even using masking to conceal information and subsequently perform predictions and analysis. The aim is to find different semantic relationships that can be described through characteristic sets, some of which may be positional, frequency-based, word occurrence, or sets of words. It is possible to perform a segmental and suprasegmental analysis of texts, which considers sound, phoneme, articulation, phonology, intonation, and pauses [3].

pp. 29–40; rec. 2024-07-12; acc. 2024-09-03

Speech segmentation comprises two distinct processes, one of them manual and the other automatic (explicit and implicit). However, the central idea and research consideration involves conducting segmentation processes automatically, in order to provide more effective processes in the search for parameters of interest. The use of these techniques has served to enable different language analysts interested in phoneme search to interpret language and its constituent parts, considering lexical characteristics.

Phonetic segmentation of languages involves locating, sectioning, and delimiting the smallest part of a word, which has an interpretation in terms of phoneme, tone, and articulation.

The phoneme, being the smallest interpretive part of a word, considers elements of importance in the analysis because it is closely linked in the vast majority of cases with other phonemes that aid in the interpretation of a word, making it difficult to separate them. The phoneme can be described as the smallest unit of sound within a language system, when phoneme is changed for another within a context, the meaning of the word also changes [11].

Phonemes are conditioned by the meaning of a statement or by a family of sounds emitted in a particular language; therefore, any given expression can be transcribed using two phonetic levels of transcription: language-dependent (psychophonic) and language-independent (physiophonic)[12].

Phonemes tend to be interpreted differently depending on the language, considering tonalities that define a word. In some cases, these tonalities can be confused with intonations; however, tone is defined as contrasts with a paradigmatic dimension. The levels of contrast can be defined in some cases as high and low, although there may be contrasts of three, four, or five levels. Thus, a tonal language can be defined as a language that has a lexically significant, contrastive tone, but relative in each syllable [6].

In Mexico, according to INALI (Instituto Nacional de Lenguas Indígenas), there is a variety of linguistic families with different tonal interpretations, such as the Otomangue family, which includes languages like Mazahua and Otomí. Currently, the proposed research for voice segmentation and analysis of indigenous languages in Mexico focuses specifically on a purely manual analysis. This involves using software such as Praat or ELAN to analyze phonetic audio files, cutting words or sentences to identify phonemes and pronunciation tonalities [8].

In this paper, we propose to perform an explicit segmentation of phonemes in the Yuhmu language (Otomí from Ixtenco, Tlaxcala, Mexico). We will use words pronounced in the language that encompass all the phonemes, along with aligned phonetic representations in writing. This will be aided by Mel-scale spectrograms and pattern search using cosine distance between the information provided by the spectrogram matrix.

2 Related Work

Language analysis aided by neural networks involves processing a vast amount of data for training and validation, considering digital audio predominantly, with transcriptions often incorporated as support. This entails aligning information between audio and text. When temporal alignment is conducted at the smallest unit of sound, the phoneme, it is referred to as speech segmentation [5].

Speech segmentation in indigenous languages of Mexico has been approached through manual analyses, utilizing digital audio representation, spectral analysis, and human perception. However, this process has been costly in terms of time and human resources. For example, [14] conducted manual segmentation using digital audio, speech signal envelope, spectrograms, and perception via Praat. Another study by [16] involved audio recordings of Náhuatl words pronounced by a single speaker, where ELAN software was used for subjective analysis of sound production, including segmentation for analysis purposes.

In [17], a description of intonational features of Hñöñhö (a variant of Otomí spoken in Tultepec, Querétaro, Mexico) is conducted through word segmentation using Praat. Praat and ELAN are considered open-source software for recording and analyzing words or phrases through: spectrograms, pitch, intensity, volume, and cochleograms using audio or video, respectively.

On the other hand, there are computational developments in automatic speech segmentation using two approaches: implicit and explicit [7]. Implicit segmentation involves supervised learning, where the system learns about the characteristics of each speech segment, then utilizes forced alignment of transcriptions using optimization. Explicit segmentation involves unsupervised learning; in some cases, it does not require forced alignment or utilizes pre-trained models for analysis processes, besides considering a reduced number of unlabeled data from pure audio [9].

An explicit segmental analysis is carried out in [9], where the boundaries of phonemes in a given utterance of Classical Arabic are identified using cosine distance similarity scores. This method achieves a total error rate of 14.48%, while the accuracy reaches 85.2% within a 10 ms alignment error.

Another segmental analysis is conducted using neural networks combined with a parameterized structured loss function, intending for the network to learn segmental representations to detect phoneme boundaries. This model utilizes the TIMIT and Buckeye corpora of English speech, achieving state-of-the-art performance in terms of F1-score and R-value, as mentioned in [10].

In [1], a phonetic segmentation analysis is conducted based on explicit (text-independent) segmentation using wavelet packet speech parametrization features and sparse representation classifier (SRC) utilizing Arabic and English (TIMIT), achieving a higher accuracy rate than the k-Nearest Neighbors (k-NN) on TIMIT.

Meanwhile, [4] proposes a single-scan method based on geometric quadrilaterals, which comprehends patterns of a speech signal. It utilizes the geometric nature of waveform trajectories, treating the input speech signal as a sequence of structural components. According to the paper, the algorithm's performance is assessed through experiments with spoken English words with an Indian native accent and Telugu sentences (an Indian language).

Phonetic segmentation is a significant problem involving the identification of boundaries of phonetic units. A challenge addressed by various researchers for over five decades. Most of the research focuses on the English language and languages with enough datasets for computational analysis. Methods like Hidden Markov Models and Deep Neural Networks require large amounts of data for training. However, there are current challenges in applying these approaches in the context of languages with limited resources [1].

This paper considers an explicit segmentation of phonemes in the Yuhmu language, which lacks a sufficient amount of data for deep learning. However, it is possible to automate segmentation processes and supervised learning. This analysis represents an initial praxis, as previously, phonetic analysis has been conducted manually through research by [2] using Praat. The intention of this process is to achieve automatic phoneme segmentation with the assistance of digital tools.

3 Yuhmu Language

Yuhmu is one of the variants of the Otomí language, spoken in Ixtenco, Tlaxcala, Mexico. It is endangered, as only a few elderly individuals (around 70 years old) maintain proficiency in its pronunciation. In some cases, individuals under the age of 60 understand the language, with no children learning it as their first language [2].

A community census by [2] indicates that there are currently around 100 speakers of the Yuhmu language, but their linguistic proficiency is not well understood. Additionally, Yuhmu lacks its own writing system, prompting efforts to represent its sounds phonetically or develop phonetic scripts.

Morphologically, Yuhmu comprises 32 phonemes represented in its pronunciations, as classified by the International Phonetic Alphabet (IPA), including 12 vowels (V) distributed between oral and nasal sounds, as shown in Fig. 1, where the expelled air exits through either the mouth or nose exclusively depending on the case. Additionally, it consists of 20 consonants (C) categorized by the location of the articulatory organs within the vocal tract. The airflow originates from the lungs, encountering obstruction from the lips or tongue. Unlike vowels, consonants may or may not be voiced, as depicted in Table 1.

For the analysis, 297 recordings of digital audio pronunciations of words were considered, encompassing all possible combinations of phonemes that form different words. Word structures in Yuhmu are generated from the following patterns: C-V, C-C-V, C-C-C-V, and C-V-V-V, these combinations may appear at the beginning, in the middle, and at the end of a word. In some cases, this structure may represent a single word. Additionally, the tone of the words is also considered, where it is possible to observe various words with high, low, and low-high tones [2].

The base dictionary used for the Yuhmu words is the one proposed by [2], which describes all the phonemes incorporated in the language and were analyzed subjectively. On the other hand, the digital recordings used have a

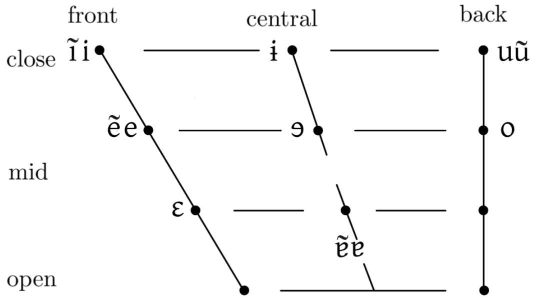

Fig. 1. Phonetic symbols of vowels (those with a tilde above them are considered nasal, while those without it are oral).

Table 1. Symbols of the International Phonetic Alphabet for Consonants in Yuhmu.

		Airway obstruction site					
Airway obstruction	Production of sound	Bilabial	Alveolar	Palatal	Velar		Glotal
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		

duration between 376 ms and 1.118 seconds, and they undergo preprocessing (noise attenuation, amplification, and word trimming).

4 Proposed Method

The following methodology is proposed to generate an explicit phonetic segmentation of the Yuhmu language, based on the analyzed literature, considering four phases (Fig. 2) described in the following sections.

4.1 Mel-scale Spectrograms

The search for optimal parameters is a fundamental aspect of research, as they contribute to improving the efficiency and effectiveness of phonetic segmentation. Therefore, we propose the parameters shown in Table 2, where spectral parameter search windowing is performed on the Mel scale. This is a technique

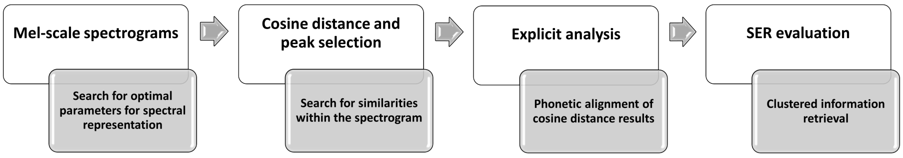
Mel-scale spectrograms
Search for optimal parameters for spectral representation
Cosine distance and peak selection
Search for similarities within the spectrogram
Explicit analysis
Phonetic alignment of cosine distance results
SER evaluation
Clustered information retrieval

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

Table 2. 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

used in signal processing to analyze short segments of an audio signal. It involves dividing the audio signal into smaller segments, called windows, which are typically overlapped to capture temporal and frequency information more precisely. Each window undergoes a Fourier analysis to convert it from the time domain to the frequency domain, allowing visualization of the signal's energy at different frequencies in that time segment[13].

Mel scale is a frequency scale perceptually relevant to humans, and it is calculated through a non-linear transformation from frequency in Hertz to the Mel scale. The conversion of standard frequency to Mel frequency can be performed using equation 1, where F represents the frequency in Hertz (Hz) of a signal. For this research, a single Hanning window is considered, as it has been previously analyzed for this language in [15]:

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

On the other hand, the window size is proposed to vary from 20 to 40 ms with a 5 ms increment to ensure a reasonable interval for analysis. The overlap is set at 25%, 50%, and 75% of this window size to minimize information loss in each analysis during the process. Furthermore, although the literature suggests using 15 to 20 Mel-filter bands (Mels), which are filters used in audio signal processing to divide the frequency spectrum into bands that mimic human auditory perception, we propose an interval of 15 to 45 Mels to capture more information within the spectrogram and its resulting matrix. Since Mels represent the rows in the matrix, a higher number of Mels will result in a greater amount of information being represented.

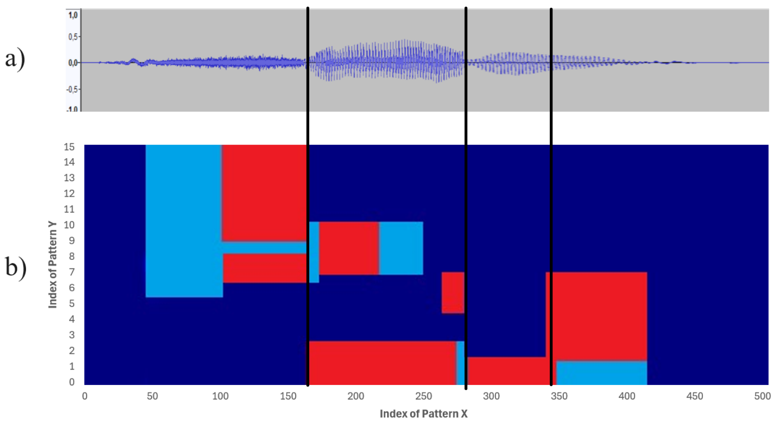

Fig. 3. a) Speech samples with phoneme boundaries over time. b) Hypothesis for ideal scores (Where navy blue tones represent null information, red tones represent high information, and light blue tones represent low information).

4.2 Cosine Distance and Peak Selection

The cosine distance (Eq. 2) is applied to the resulting spectrograms from the previous analysis. The purpose of this is to observe how similar the columns of the resulting spectrogram matrix are and to group information, separating the data corresponding to each of the phonemes. Ideally, we aim to establish a contrast of information (Fig. 3), considering the digital representation of audio and the segmentation generated from it. The cosine distance is defined as follows:

$$\text{cosine\_distance}(x, y) = \frac{\sum_{i=1}^{n}(x_i)^2 \cdot \sum_{i=1}^{n}(y_i)^2}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i)^2} \cdot \sum_{i=1}^{n}(x_i \cdot y_i)}, \quad (2)$$

where:

x_i and y_i are the components of vectors $\mathbf{x}$ and $\mathbf{y}$ respectively.

n is the length of the vectors (the number of components).

The symbol $\cdot$ represents the dot product between two vectors.

Following the cosine distance calculation between columns of the spectral matrix, a search for relevant information is conducted for each resulting matrix. Regardless of the parameter combination, the aim is to obtain the top 25% scoring points in each column, considering a threshold similar to the third quartile, which varies in each case, taking into account the number of Mels and windowing.

4.3 Explicit Analysis and SER Evaluation

Within the explicit analysis, a manual search for grouping information within the resulting matrices is considered. Segments found for each word are observed,

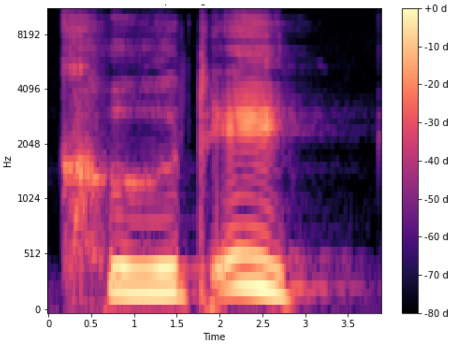

Fig. 4. Mel-scale spectrogram of the word 'ceñidor' in Yuhmu.

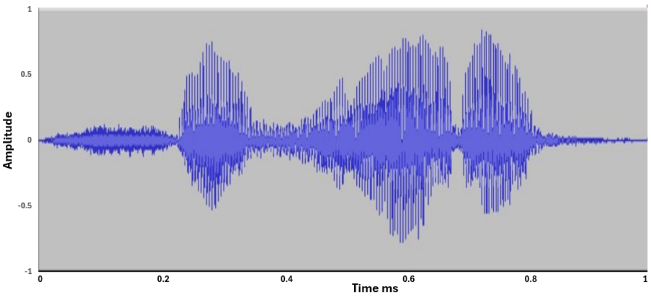

Fig. 5. Digital representation of an audio signal of the word "ceñidor" in Yuhmu.

the sample is representative, and delimited segments can be observed, recording the resulting number of segments in a table. Then, a comparison is made with the phonemes that each word actually contains. This comparison will be named Evaluation SER (Segment Error Rate), statistically indicating the segments generated by the proposed method and the percentage of error.

5 Experimental Results

When performing the methodological processes outlined in Section 4, we can obtain three matrices that are represented in figures, which describe the behavior of the information grouped in each of these. The main idea is to eliminate the connections that may link a phonetic representation to numerical information, considering three processes. The first one analyzes a Mel-scale spectrogram (see Fig. 4), which has as a precedent a digital audio signal represented in samples over time (see Fig. 5).

Having the representation in Mel-scale, cosine distance is applied between two vectors corresponding to the columns of the Mel-scale spectrogram matrix.

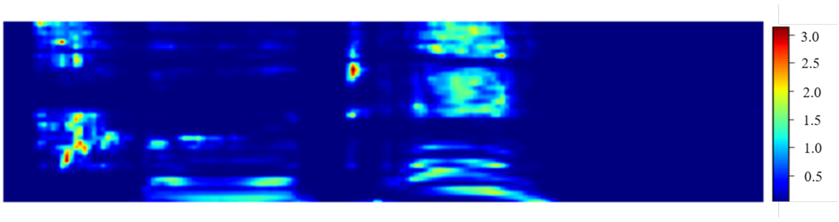

Fig. 6. Resultant matrix from cosine distance calculation of the word 'ceñidor' in Yuhmu.

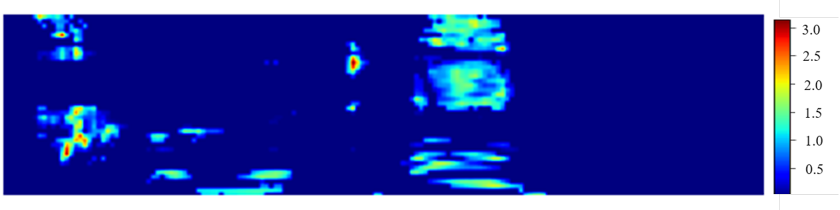

Fig. 7. Resultant matrix of maximum values per column of the word 'ceñidor' in Yuhmu.

These represent the spectral energy distribution of an audio segment based on the bands (for our case, the energy of the phonemes of the words in Yuhmu). Each column of the Mel-scale spectrogram shows how the spectral energy varies over time for a specific Mel frequency band. By calculating the cosine distance, scores are generated that provide a relationship between the values of each column. This will help us to create greater contrasts within the columns, generating high contrasts if they are dissimilar and low contrasts if they are similar, producing a separation in terms of energy and visualization of the spectrogram (see Fig. 6).

The process to obtain the search for contrasts still considers some connection between different groups of information representing the phoneme of a word. Thus, a selection of information from matrices is used, similar to the use of quartiles to obtain a threshold. However, in this case, it is obtained from the number of rows of the matrix, approximately 25% of high data for each column, which generates greater contrast between phonetic separation in the vast majority of words (see Fig. 7).

Finally, upon segmenting the data within the matrix, results of phonetic segmentation are obtained. Table 3 presents comparisons of the segmentation performed and the actual number of phonemes comprising a word. Ultimately, an average SER of 23.89% to 26.03% is obtained, varying among the combinations of windowing and the quantity of Mels used, which are represented in rows within the spectral representation (if the resulting spectral matrix was calculated with

Table 3. Comparison of some phonetic representations found in the spectrogram and actual phonemes, along with the resulting SER value.

Word 'yuhmu' represented in Spanish	Quantity of phonemes	20 ms (Windowing) 25 % (Overlap) 15 (Mels)	Difference between phonemes	SER %	20 ms (Windowing) 25 % (Overlap) 20 (Mels)	Difference between phonemes	SER %	20 ms (Windowing) 25 % (Overlap) 25 (Mels)	Difference between phonemes	SER %
abeja	4	4	0	0	4	0	0	4	0	0
abuela	3	2	1	33.33	2	1	33.33	2	1	33.33
abuelo	4	3	1	25	3	1	25	3	1	25
frijol	6	5	1	16.66	5	1	16.66	5	1	16.66
gallina	4	3	1	25	3	1	25	3	1	25
grueso	4	2	2	50	2	2	50	2	2	50
huazontle	6	4	2	33.33	4	2	33.33	4	2	33.33
niño	8	7	1	12.5	7	1	12.5	7	1	12.5
sal	1	1	0	0	1	1	0	1	1	0
viento	5	4	1	20	4	1	20	4	1	20

15 Mels, it will have 15 rows), providing energy information over time within the Mel scale.

The transformation from the Mel-scale spectrogram to the matrix selecting peaks considers groupings and contrast of information levels, achieving differences between the mentioned aspects regarding each phoneme. It is important to highlight that the results obtained from the average SER for each combination of spectral representation parameters of the words yield better results with a window size of 20 ms and a 25% overlap, along with a high number of Mels, ranging between 35-45. This consideration takes into account the methodological section, where a higher number of Mels implies a greater number of rows in the matrix, and therefore, more information in the final matrices. However, it does not represent a significant difference compared to the results with 15 to 35 filters.

Perhaps visually significant changes between Fig. 6 and Fig. 7 may not be perceived; however, there are datasets that are eliminated, generating a greater contrast between the datasets. At the matrix level, the given data are functional for performing groupings of these datasets, considering that we are not working at the image level but at the level of matrix information.

On the other hand, the SER considered for this experiment encompasses significant resultant values, as the literature cited displays segmentations with error rates ranging from 5-20%. However, these considerations involve languages with ample data for training and validating artificial neural networks. Taking this into account, the method presented yielded favorable results for a language with limited digital resources, encompassing all phonetic characteristics of the language.

6 Conclusions

Phonetic segmentation in indigenous languages of Mexico represents a significant challenge due to the linguistic and phonetic diversity of these languages. The results of our study indicate that the application of digital signal processing techniques, such as Mel-scale spectrogram analysis and cosine distance between vectors, can provide an effective approach for phonetic segmentation in these languages. This is particularly relevant given the importance of preserving and adequately analyzing the phonetic structures of indigenous languages for linguistic purposes.

Comparing the results of phonetic segmentation with the actual number of phonemes in a word provides valuable information about the accuracy and effectiveness of the techniques used. This comparative analysis reveals both successes and potential areas for improvement in the segmentation process, which can guide future research efforts and methodological refinements in this field, ultimately proposing feedback on mispronunciation aided by its comparative and location of the mispronounced phoneme.

Phonetic segmentation in indigenous languages of Mexico is an evolving field that requires the application of interdisciplinary approaches. With the improvement of techniques and understanding of the unique phonetic characteristics of each indigenous language, it is possible to advance towards better documentation and preservation of these important cultural and linguistic expressions.

As a future work, it is planned to analyze a language with similar phonetic characteristics such as Jñatrjo (Mazahua from the State of Mexico), which belongs to the same linguistic family as Otomanguean languages and shares some similar words. However, Jñatrjo is distinct in having a developed system of writing and phonetic representation of words.

This segmental analysis aims to identify both good and poor pronunciation of words in both languages using a phonetic approach. This could significantly contribute to improving the understanding and preservation of the accurate phonetic structures of these indigenous languages, facilitating their proper documentation and cultural transmission.

Acknowledgments. This work is supported by CONAHCYT scholarship number 1169943, BUAP-VIEP 189 and UPIIT-IPN 20242590 research projects.

References

1. Al-Hassani, I., Al-Dakkak, O., Assami, A.: Phonetic segmentation using a wavelet-based speech cepstral features and sparse representation classifier. Journal of Telecommunications and Information Technology (4), 12–22 (2021)
2. Alarcon Montero, R.: Manual para la escritura de los sonidos del yuhmu. INAH (2023)
3. Ávila, J., Díaz, T., Ávila, C., Concepción, C., Carmona, E., Robaina, C., Cárdenas, C., Hernández, C., Mederos, P., Rouco, E.: Didáctica de la lengua española I. Editorial Pueblo y Educación (2013), https://books.google.com.mx/books?id=RR6owwEACAAJ
4. Bhagath, P., Das, P.K.: Quadrilaterals based phoneme segmentation technique for low resource spoken languages. In: TENCON 2022-2022 IEEE Region 10 Conference (TENCON). pp. 1–6. IEEE (2022)

5. Brognaux, S., Drugman, T.: Hmm-based speech segmentation: Improvements of fully automatic approaches. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24(1), 5–15 (2015)
6. Gussenhoven, C.: The phonology of tone and intonation (2004)
7. van Hemert, J.P.: Automatic segmentation of speech. IEEE Transactions on Signal Processing 39(4), 1008–1012 (1991)
8. INALI: Catalogo de las lenguas indígenas nacionales. Retrieved in October 2022 from https://www.inali.gob.mx/clin-inali/ (Fecha de acceso: octubre de 2022)
9. Javed, M., Baig, M.M.A., Qazi, S.A.: Unsupervised phonetic segmentation of classical arabic speech using forward and inverse characteristics of the vocal tract. Arabian Journal for Science and Engineering 45, 1581–1597 (2020)
10. Kreuk, F., Sheena, Y., Keshet, J., Adi, Y.: Phoneme boundary detection using learnable segmental features. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8089–8093. IEEE (2020)
11. Kyriakopoulos, K.: Deep learning for automatic assessment and feedback of spoken english. Ph.D. thesis (2022)
12. Moore, R.K., Skidmore, L.: On the use/misuse of the term'phoneme'. INTERSPEECH (2019)
13. Pangaonkar, S., Panat, A.: 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)
14. Penner, K.: Prosodic structure in ixtayutla mixtec: Evidence for the foot (2019)
15. Ramos-Aguilar, E., Olvera-López, J.A., Olmos-Pineda, I.: A general overview of language pronunciation analysis based on machine learning. Res Comput Sci 152 (2023)
16. Turnbull, R.: The phonetics and phonology of lexical prosody in san jerónimo acazulco otomi. Journal of the International Phonetic Association 47(3), 251–282 (2017)
17. Velásquez Upegui, E.P.: Entonación del español en contacto con el otomí de san ildefonso tultepec: enunciados declarativos e interrogativos absolutos. Anuario de letras. Lingüística y filología 8(2), 143–168 (2020)