ISSN 2007-9737

WavLM-Based Automatic Pronunciation Assessment for Yuhmu Speech: A Low-Resource Language

Eric Ramos-Aguilar[1,2], J. Arturo Olvera-López[1,*], Ivan Olmos-Pineda[1]

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

[2] Instituto Politécnico Nacional, UPIITA,
Mexico

eric.ramosag@alumno.buap.mx, {jose.olvera, ivan.olmos}@correo.buap.mx

Abstract. This paper presents an approach to classify correct and incorrect pronunciation in Yuhmu, an endangered Indigenous Minority Language, using acoustic embeddings combined with SVM and MLP models. Unlike typical low-resource language tasks focused on automatic speech recognition (ASR) or machine translation, this work employs deep acoustic representations to detect phonetic quality, achieving high accuracy and consistency across different embedding sizes. The results highlight the potential of leveraging labeled audio data and advanced speech models like WavLM to provide phonetic feedback and support language revitalization. This research establishes a foundation for deeper computational phonetic analysis in Yuhmu and opens avenues for future exploration in direct audio-to-audio translation, automatic phonetic segmentation, and detailed phoneme-level evaluation, contributing to the documentation and preservation of underrepresented languages.

Keywords. Low resource languages, Yuhmu language, supervised learning, speech analysis.

1 Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence that aims to endow computational systems with the ability to automatically process human language, including its comprehension, generation, and structural transformation [19]. Within this framework, the automatic analysis of pronunciation emerges as a critical component, operationalized through models that capture acoustic-phonetic representations based on articulatory parameters (vocal tract position, voicing) and perceptual cues (formant frequencies, spectral envelope) [22].

Pronunciation comprises two essential skills: oral production (speaking) and auditory comprehension (listening). Speaking involves the practical application of the phonological features of the target language (TL), at both the segmental level (phonemes and their combinations) and the suprasegmental level (stress, rhythm, and intonation), requiring conscious articulation for effective communication. Conversely, listening demands the decoding and interpretation of these same elements in others' speech, allowing the learner to recognize and assimilate the phonological features of the TL.

According to [22], the simultaneous development of both skills is crucial for pronunciation acquisition, as their interaction reinforces phonological competence. While speaking emphasizes the active execution of sounds, listening facilitates their internalization through auditory exposure, thus establishing a dynamic cycle between production

and perception that supports comprehensive language learning.

The analysis of suprasegmental elements represents a significant methodological challenge in phonetic research, particularly when evaluating isolated words. This phenomenon creates an epistemological dilemma between two approaches: (a) the precise assessment of segmental and suprasegmental pronunciation, and (b) the evaluation of overall speech intelligibility. As noted by [1], this duality compels researchers to establish specific diagnostic criteria for pronunciation while also providing meaningful feedback on the analyzed lexical units.

Computer-Assisted Pronunciation Training (CAPT) systems employ various evaluation metrics to analyze oral production through automated scoring methods. These metrics operate at different levels of linguistic analysis (complete utterances, isolated words, or specific phonemes), enabling a granular assessment of the speaker's phonetic performance.

A critical aspect in the development of such systems lies in the distinction between High-Resource Languages and Low-Resource Languages (LRLs). This classification reflects the differing data requirements of machine learning models, which demand substantially different volumes of training data to achieve accurate pronunciation analysis. While high-resource languages typically possess extensive phonetic corpora for development, LRLs face significant technical challenges due to the scarcity of available data, limiting the effectiveness of CAPT systems in these linguistic contexts.

The computational study of Indigenous Languages faces significant challenges due to their status as LRLs. Unlike high-resource languages with extensive digital corpora, exhaustive grammatical descriptions, and well-developed computational tools, these languages lack sufficient documentary resources to capture their complex dialectal variability. This limitation manifests in the absence of systematic digitized materials, hindering both linguistic analysis and the development of specialized technological applications.

This issue is exacerbated by two concurrent factors: first, the phonetic and grammatical particularities of these languages, which present greater intra-linguistic variation; and second, the challenges in their intergenerational transmission, where phenomena such as disfluency and resistance to learning compromise their linguistic vitality. This situation not only complicates their academic study but also threatens their very preservation, requiring adaptive computational approaches capable of operating with limited data while documenting the structural richness of these vulnerable languages.

In Mexico, there are sixty-eight Indigenous Languages distributed throughout the national territory, with a higher concentration in the southern and central regions of the country. These languages exhibit significant internal diversity, as the linguistic variants spoken in different communities give rise to unique regional forms. According to the Indigenous Languages Center [11], these languages are grouped into eleven linguistic families: Álgica, Yuto-Nahua, Cochimi-Yumana, Seri, Otomangue, Maya, Totonaco-Tepehua, Tarasca, Mixe-Zoque, Chontal and Huave.

The analysis of Mexican Indigenous Languages (MIL), as well as other low-resource languages, faces multiple challenges due to the limited availability of audio and text data. To address these limitations, a common strategy is transfer learning, which enables the evaluation of these languages using data from other languages. Additionally, techniques such as the use of acoustic and phonetic embeddings derived from high-resource languages, or the training of neural networks based on similar phonemes, are often employed.

To perform pronunciation evaluation processes in MIL, various researchers primarily linguists have collected recordings of words or sentences spoken by native speakers from diverse communities. These efforts have facilitated the construction of digital audio corpora, typically involving between five and thirty participants. An example is provided by [17], which documents a corpus consisting of six native speakers of Mixteco, specifically of the still underdescribed variety spoken in Ixtayutla, Oaxaca, which belongs to the Oto-Manguean language family.

This study presents an analysis for the classification of correct and incorrect pronunciation in the Yuhmu language at the segmental level.

The research establishes the methodological foundations to examine the phonetic components of words in this language through acoustic embeddings extracted from two WavLM-based models. Section 2 analyzes previous studies related to the processing of LRL, with an emphasis on MIL. Section 3 details the database used for the analysis, while Section 4 describes the proposed methodology. Subsequently, Section 5 presents the results obtained. Section 6 provides an analysis of the results by comparing them with the existing literature, and finally, Section 7 discusses the study's conclusions and outlines directions for future work.

2 Related Work

Pronunciation analysis considers relevant elements in the phonetic composition of words. Methods used to classify pronunciation features were initially developed via Computer-Assisted Pronunciation Training (CAPT) systems based on Hidden Markov Models (HMM), which have been implemented in some LRL for various types of analysis.

A comprehensive review of LRL [14] examined progress in the processing of African, Indian, Turkic, and Niger-Congo languages, highlighting techniques such as automatic corpus projection and alignment (via distributional similarity and lexical induction) to transfer annotations from high-resource languages, as well as key models such as Multilayer Perceptrons (MLP) and HMMs in Automatic Speech Recognition (ASR) adapted to LRL phonetic units, and multilingual embeddings for machine translation. Results indicate significant advances (a F1 metric value of 0.44 in Part-of-Speech Tagging).

In another study, [8] considered the analysis of LRLs where a speech recognition system for low-resource African languages such as Maninka, Susu, and Pular was developed using radio recordings to train unsupervised models. A custom encoder named West African wav2vec was built, achieving performance comparable to or better than larger commercial models with less data. This system enables the creation of a virtual assistant capable of recognizing voice commands with high accuracy (up to 88.1%), facilitating digital access for illiterate populations in West Africa.

In [25], the development of a part-of-speech tagger for the Khasi language, an Austroasiatic language spoken in northeast India, is addressed. Khasi lacks digital linguistic resources such as annotated corpora or natural language processing tools. To overcome this limitation, the authors constructed a corpus comprising approximately 103,998 words and applied Brill's transformation-based learning method for automatic part-of-speech tagging. This approach achieved an accuracy of 97.73% on the validation set and 95.52% on the test set, demonstrating the effectiveness of the method in a low-resource setting.

In [6], the XLSR model was evaluated on LRLs. Within the CommonVoice corpus, languages such as Swedish (3 hours), Turkish (11 hours), and Tatar (17 hours) are notably low in data. Similarly, the BABEL dataset includes Swahili (30 hours) and Tok Pisin (36 hours). Results show that XLSR achieves significant improvements in these languages, such as a 67% reduction in phoneme error rate for Swedish, thanks to multilingual transfer learning from high-resource languages.

Another study applied transfer learning to enhance speech recognition in Amharic, an LRL, by adapting pretrained English and Mandarin models. The English-based model reduced word error rate (WER) from 38.7% to 24.5%, outperforming the Mandarin model (28.5%). These results demonstrate the effectiveness of this approach for languages with data scarcity, particularly when knowledge is transferred from dominant languages such as English [26].

Additionally, research efforts have addressed improvements in ASR for LRL using multilingual models, self-supervised learning, and cross-lingual adaptation. For instance, [2] proposed an ASR corpus for Quechua based on the Siminchik dataset and evaluated SSL models such as XLSR-53, XLS-R 128, and mHuBERT on six Indigenous American languages, finding that XLS-R 128 performed best, with an average Character Error Rate (CER) of 36.8%. [20] introduced wav2vec, a convolutional SSL model trained without labels, reducing WER by up to 36% in low-resource English settings.

The work proposed in [12] applied wav2vec 2.0 combined with automatic language identification on Indic languages such as Tamil and Malayalam, achieving significant improvements in language detection accuracy and ASR adjustment. In [23] a multilingual ASR systems are developed using the GlobalPhone corpus and adapted models to four Ethiopian languages (Amharic, Oromo, Tigrinya, Wolaytta), observing a WER reduction of up to 51.41% when incorporating related language data.

Finally, [7] addressed educational inequality through machine learning models applied to PISA-D data, showing that Indigenous youth in Guatemala, Paraguay, and Senegal are 13% to 20% less likely to reach basic proficiency levels in reading and mathematics.

The development of LRL processing has had a significant impact across various regions of the world, particularly in African and Asian languages, as previously discussed. In contrast, the analysis of Latin American Indigenous languages especially those of Mexico has progressed more slowly in the context of machine learning, mainly due to the lack of structured and computable digital data.

Several recent studies have tackled machine translation (MT) for Indigenous American languages, with particular focus on those spoken in Mexico. In [28], 10 Indigenous languages were studied, including five from Mexico: Nahuatl, Hñähñu (Otomi), Wixarika (Huichol), Rarámuri, and Bribri. A multilingual system was trained using mBART pretraining with 13 GB of monolingual data from high-resource languages and fine-tuned with 140 MB of parallel data. The system showed improvements over the baseline with an average increase of 1.64 BLEU and 0.0749 CHRF; for instance, Nahuatl achieved BLEU = 1.2 and CHRF = 0.238, while Wixarika reached BLEU = 6.74 and CHRF = 0.229. BLEU measures n-gram overlap between machine translation and reference, evaluating word-level precision, while CHRF assesses similarity using character n-grams, being more sensitive to morphological and orthographic variations.

Later, [9] reported the results of the 2023 shared task involving 11 languages (six from Mexico), including a new professional corpus for Chatino. The best-performing system (Sheffield) showed an average CHRF improvement of +9.64 points compared to 2021. Highlights include Nahuatl (CHRF = 27.25), Otomi (CHRF = 15.30), and Chatino (CHRF = 39.97).

Although the focus of [27] was on Formosan languages, they also applied their technique to the Spanish–Nahuatl pair using a parallel corpus of 16,145 sentences and a bilingual lexicon from AULEX. Three strategies were evaluated (lexicon as parallel data, pseudo-parallel data, and a combination of both), with the best result achieved by the latter: +5.55 BLEU and +10.33 CHRF for Spanish–Nahuatl. These studies demonstrate substantial progress in MT quality for Indigenous languages, with accumulated improvements of up to +5.55 BLEU and +10.33 CHRF in the case of Nahuatl.

In [5], the Indigenous Mexican languages Nahuatl and Wixarika were examined. MT models were trained using multilingual transfer techniques (mBART50 and mBART50curr), and their performance was evaluated on Spanish translation tasks. The results were promising, with BLEU scores reaching up to 12.74 for Nahuatl and 7.84 for Wixarika. The study demonstrated that curriculum transfer strategies can significantly enhance performance, even without artificially augmented resources.

The work by [21] focused on Highland Puebla Nahuatl, a low-resource language of the Uto-Aztecan family. The study introduced an open speech translation corpus to document this endangered language. Speech translation (ST) models were compared, including cascaded systems (ASR + MT) and end-to-end models, with the latter outperforming the former. Results showed that end-to-end approaches are promising for resource-scarce languages.

Another study [13] explored MT for Nahuatl, Otomi, and other Indigenous American languages. Both statistical and neural models were employed, and the use of external data was allowed. Results demonstrated that multilingual models and normalization techniques significantly improved translation quality.

In turn, [16] conducted a MT analysis including the Nahuatl–Spanish pair. Statistical (IBM Model 2) and neural (transformers) models were evaluated

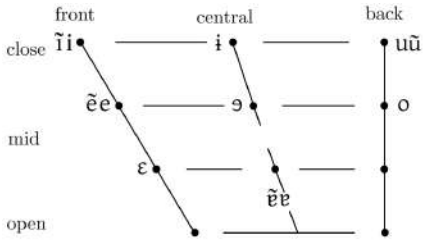

Fig. 1. Phonetic symbols of vowels (those with a tilde above them are considered nasal, while those without it are oral)

under one-to-one and one-to-many configurations. The results indicated that statistical models remain competitive for LRLs, although neural models yielded improvements in some cases. Second-best results were achieved for Spanish–Bribri and Spanish–Ashaninka pairs, underscoring the importance of adapting approaches to the characteristics of each language.

Finally, [24] provided a broad overview of NLP advances for Latin American Indigenous languages, including Mexican languages such as Nahuatl, Otomi, Mazateco, Tepehua, Mixteco, and Popoluca. It was found that out of the 68 recognized Indigenous languages in Mexico, only 22 have been explored within NLP. Key tasks addressed include MT, morphological analysis, named entity recognition, and ASR. Nahuatl and Otomi stood out for having multiple tasks developed. Overall, approximately 40% of the research focused on machine translation. Surveys were conducted with over 350 speakers and 27 researchers to identify challenges and opportunities.

While the analysis of MIL has shown promising progress, their development has been mainly concentrated in areas such as translation and morphological analysis, with limited attention to tasks related to pronunciation and speech processing. Despite advances in multilingual and transfer learning techniques, these have not yet been widely applied to the phonetic analysis of these languages.

Although the analysis of Mexican Indigenous Languages has shown promising progress, their development has primarily occurred in areas such as translation and text-based work. However, the use of audio data and parallel corpora for their analysis still represents a significant challenge within the research field, even in studies that lay the foundational groundwork for their investigation.

For this reason, the present work specifically focuses on the analysis and classification of correct and incorrect pronunciation in the Yuhmu language from digital audio recordings, a variant of Otomi. To achieve this, characteristic embeddings are employed that allow the identification of patterns based on semantic relationships within the audio data.

3 Yuhmu Language

Yuhmu language is a variant of Otomi (a macrolanguage within the Otomangue linguistic group, spoken by an ethnic and cultural group distributed across the south-central region of Mexico), specifically 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 of age understand the language, but there are no children acquiring Yuhmu as their mother tongue [10], which has caused a decline in the language and places it at risk of extinction. Consequently, it is considered a Low-Resource Language due to the scarcity of digital data available for computational analysis.

According to a community census conducted by [10], there are approximately ± 75 Yuhmu speakers, although their level of linguistic competence has not been thoroughly documented. The language lacks a native writing system, which has led to efforts aimed at phonetically representing its sounds through the development of isolated writing systems proposed by various historians.

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

Table 1 summarizes the phonetic representation of Yuhmu consonants. This table contains two

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

		Airway obstruction site					
Airway obstruction mode		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		

main columns, each describing more specific features:

- *Manner of airflow obstruction*: Describes how the airflow is modified as it passes through the vocal tract, using terms such as plosive, affricate, fricative, nasal, among others, and whether the sounds are voiced or voiceless (i.e., whether or not vocal fold vibration occurs).

- *Place of airflow obstruction*: The specific location within the respiratory tract where the airflow is obstructed, impeding the normal passage of air from the lungs. Depending on the position of the tongue, lips, or glottis, the obstruction may be classified as bilabial, alveolar, palatal, velar, or glottal.

Given that Yuhmu, like many Indigenous languages, lacks a standardized writing system, its transmission and preservation rely fundamentally on orality. This makes phonetic and acoustic analysis particularly relevant, as sounds constitute not only the primary means of communication but also the main medium for linguistic documentation. In this context, considering the absence of a conventional writing system and the sole availability of phonetic representations, the use of digital audio is proposed as a form of analysis.

Our audio sample consists of 24 Yuhmu speakers (12 native speakers, 7 non-native speakers with good pronunciation, and 5 non-native speakers with poor pronunciation), representing approximately 32% of an estimated total population.

This proportion is appropriate for small populations and is especially important in the study of Low-Resource Languages [18], as such languages typically have fewer speakers and limited digital resources for their documentation.

The digital audio dataset consists of 5,835 correct pronunciations and 2,620 incorrect pronunciations, based on a core dictionary of 330 words that cover all phonemes of the language for machine learning analysis. The words include phonetic combinations that form different words in Yuhmu. The word structure in Yuhmu is generated from the following patterns: C-V, C-C-V, C-C-C-V, and C-V-V-V. These combinations can occur at the beginning, middle, or end of a word, and in some cases, represent a single word. The tonal aspect of the words is also considered, observing variations of high, low, and low-high tones [10].

The base dictionary used for Yuhmu words is the one proposed by [10], which describes all phonemes incorporated in the language and is subjectively analyzed. The digital recordings vary in duration from 376 ms to 1.118 s. The recordings include a transcription of the phonemes present in each word, which does not follow a conventional writing system but rather represents 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 accessible upon reasonable request under specific conditions.

4 Proposed Methodology

The proposed methodology for classifying correct and incorrect Yuhmu pronunciation consists of three stages, as illustrated in Figure 2.

First, data preprocessing is performed, where the analysis of digital audio information involves data cleaning and normalization.

In the second stage, two pre-trained WavLM models are used to obtain embeddings. Finally, in the third stage, two classifiers are considered, and through hyperparameter tuning using grid search, the best model for classifying the Yuhmu language is identified.

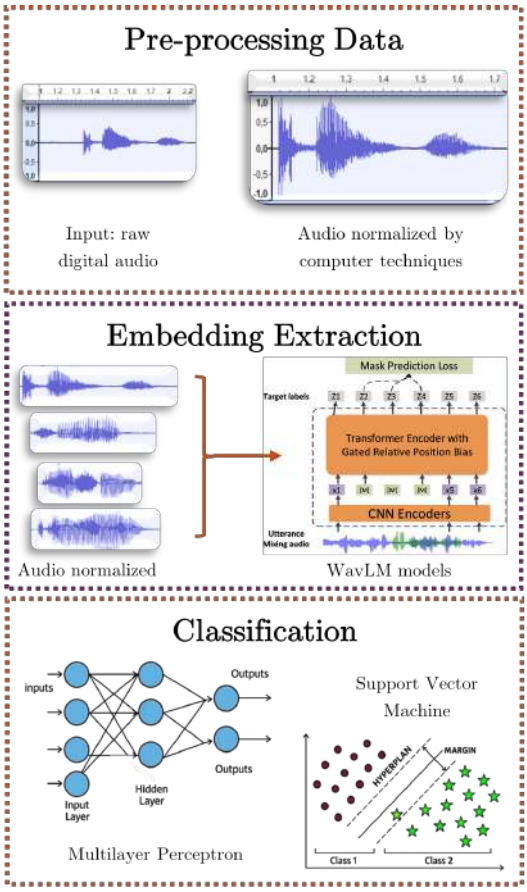

Fig. 2. Methodology phases for classification of Yuhmu language

4.1 Pre-processing

The use of computational techniques has been instrumental in the development of data pre-processing. The audio recordings, which are integrated in the form of sentences, tend to contain background noise; therefore, it is necessary to perform digital audio segmentation. This is essential because the analysis in this research is conducted at the word level, in addition to removing elements irrelevant to pronunciation analysis.

For preprocessing the digital audio, the open-source software Audacity is used, which allows audio editing and digital sound recording of native and non-native pronunciations of the language. The audio files follow a standardized normalization procedure:

1. Recordings are unified into a single channel and converted to WAV format (Waveform Audio Format), which enables uncompressed recording, ensures compatibility with software devices, and allows storage of additional metadata. The use of a single channel facilitates analysis in computational and closed environments.
2. Background noise or signals unrelated to the words in the audio segments are identified and attenuated.
3. The segmented digital audio signals corresponding to individual words are amplified to 12 dB, nearly doubling the auditory power, which improves the audio volume and enhances perception of the words' sounds.
4. If necessary, noise attenuation is performed again, since amplification tends to increase unwanted signals such as electrical noise.
5. Finally, the audios are imported and labeled according to correct or incorrect pronunciation.

4.2 Embedding Extraction

Considering the preprocessed digital audio, it is possible to begin analyzing important signal features, which may be in the time or frequency domain, taking into account various aspects referenced in the literature for analyzing audio signals representing speech.

WavLM is a pre-trained model proposed by [4] designed to address a variety of speech processing tasks. This model jointly learns masked speech prediction and noise reduction during pre-training.

This means that WavLM not only preserves the ability to model speech content through masked speech prediction but also enhances its potential for non-automatic speech recognition tasks by improving speech noise reduction.

The model is trained on a large dataset, ranging from 60,000 to 94,000 hours of data. Its architecture is based on the Transformer model,

Table 2. Sample distribution by gender and pronunciation type, where h denotes male and m denotes female

gender	Correct natives		Correct non-natives		Incorrect non-natives	
	train	test	train	test	train	test
h	3	2	3	1	2	1
m	5	2	2	1	1	1

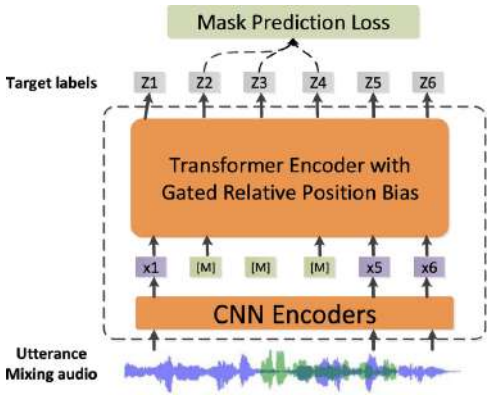

Fig. 3. Model Architecture of WavLM

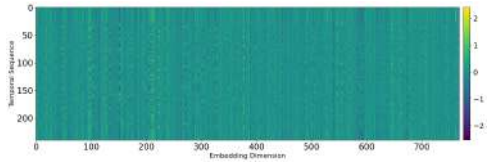

Fig. 4. Representation of proposed embedding as a characteristic tensor

comprising a convolutional feature encoder and a Transformer encoder (Fig. 3).

The proposal also includes a simulation of noisy/overlapping speech with multiple speakers and various background noises for self-supervised pre-training.

Speech classification has proven useful in majority languages, serving as a fundamental basis for analyzing whether a word is pronounced correctly or not. For the present methodological development, two pre-trained models are considered that output embeddings (numerical representations of elements such as words or phrases in a vector space that capture their meanings and semantic

Table 3. Model comparison: overall performance and consistency in pronunciation analysis by gender

Model	Accuracy	Precision	Recall	F1-score	Std. Dev.	Embedding size
SVM-1	0.8586	0.8322	0.8855	0.8770	0.0204	256
MLP-1	0.8441	0.8895	0.8333	0.8581	0.0211	256
SVM-2	0.8381	0.8386	0.9479	0.8889	0.0451	768
MLP-2	0.8402	0.8621	0.9157	0.8877	0.0282	768

relationships [3]), with the aim of performing classification through embeddings.

The first model is "WavLM-Base-Plus for Speaker Diarization" [15], which provides embeddings of size 512 and focuses on voice diarization the process of identifying and separating voices of different speakers within a recording. The second model is the base "WavLM" [4], which outputs a tensor of embeddings sized 1×N×768, where N is the temporal dimension of the analyzed audio. This can be visualized as shown in Fig. 4, where the X-axis (Embedding Dimension) represents the dimensionality of each generated embedding, and the Y-axis (Temporal Sequence), which segments the temporal duration of each digital audio signal into 20 ms intervals, varies according to the duration of the analyzed audio segment.

The colors correspond to the different numerical representations of the resulting embeddings. It is important to note that these values do not have physical units, as embeddings are dimensionless and constitute an abstract numerical encoding of the relevant features of the input data, as well as their semantic relationships.

To maintain a consistent analysis standard, it was proposed to convert the embedding tensor into a single vector representation, similar to the representation provided by the other model. Therefore, an average was computed across the columns of the resulting tensors to obtain an embedding of size 768. Thus, there are two embedding representations from the two models, with sizes 512 and 768 respectively.

4.3 Classification

The representation of audio through embeddings enables the establishment of semantic relationships in NLP. To visualize these representations in

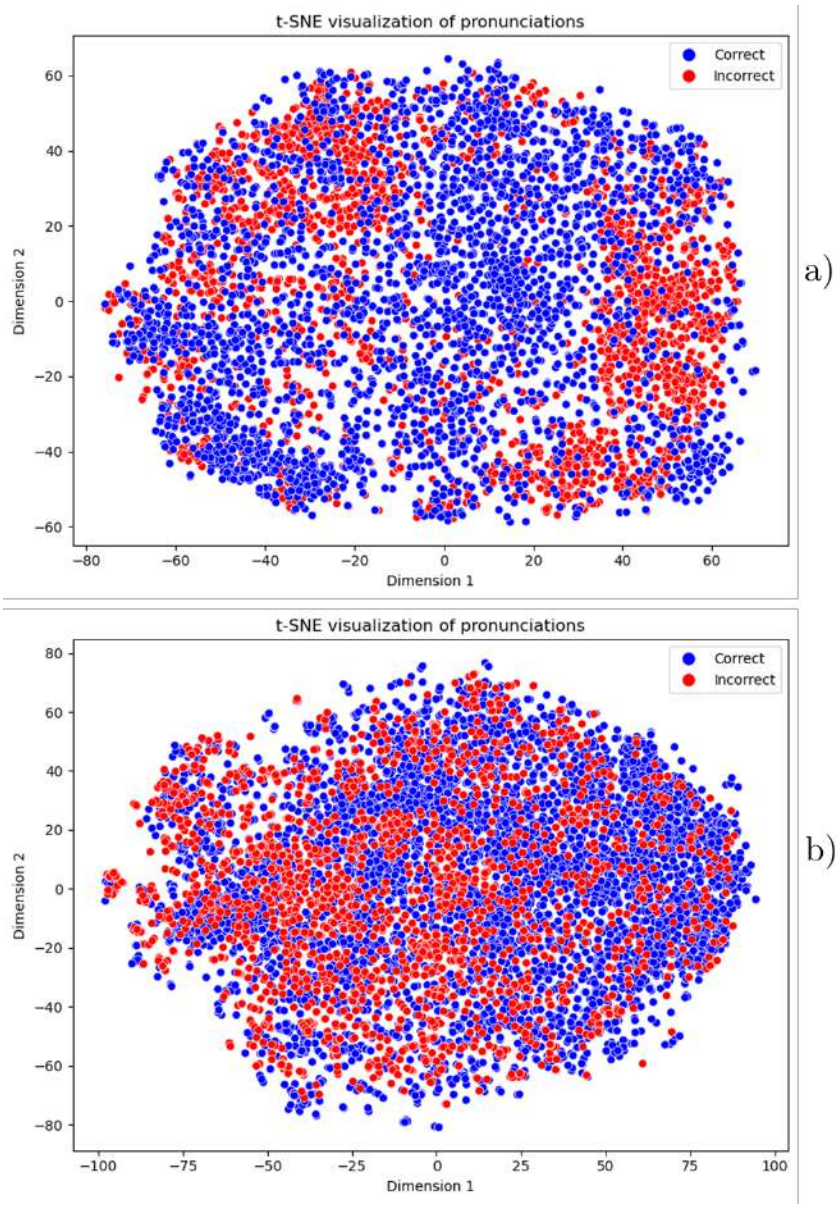

Fig. 5. Visual representation of the dataset pronunciation embeddings: a) 256-dimensional and b) 768-dimensional

an interpretable manner, the t-SNE (t-distributed Stochastic Neighbor Embedding) technique has been used. t-SNE is a nonlinear dimensionality reduction method primarily employed for the visualization of high-dimensional data, projecting embeddings into a two-dimensional space. In this visualization, two groups of embeddings have been plotted, facilitating the identification of patterns and differences between both data sets (Fig. 5).

Following the visualization of embeddings, two classification models were trained: a Multilayer Perceptron (MLP, a neural network composed of an input layer, one or more hidden layers, and an output layer, where each neuron applies a

Table 4. Comparison of previous works on Low-Resource Languages (LRL). (ASR: Automatic Speech Recognition, MT: Machine Translation, ST: Speech Translation)

Reference	Language(s)	Task	Main Technique	Highlighted Result
Doumbouya et al. (2021)	Maninka, Susu, Pular	ASR	Regional wav2vec	Accuracy: 88.1% (commands)
Conneau et al. (2020)	Swedish, Turkish, Tatar	ASR	XLSR	67% reduction in PER
Woldemariam et al. (2020)	Amharic	ASR	Transfer learning from English	WER: 38.7% $\rightarrow$ 24.5%
Chen et al. (2023)	Quechua and 5 others	ASR	XLS-R 128 (SSL)	Average CER: 36.8%
Zheng et al. (2021)	Nahuatl, Otomi, etc.	MT	Multilingual mBART	BLEU: up to 6.74; CHRF: 0.238
Ebrahimi et al. (2023)	Chatino, Otomi, etc.	MT	mBART + professional corpus	CHRF: Otomi 15.30; Chatino 39.97
Chen et al. (2022)	Nahuatl, Wixarika	MT	mBART50curr	BLEU: up to 12.74
Shi et al. (2021)	Nahuatl (Puebla)	ST (speech to text)	ASR + MT / Direct ST	ST outperforms cascade model
Tonja et al. (2024)	Nahuatl, Otomi, etc.	Multitask NLP	Systematic review	40% focused on MT (22/68 languages studied)
Present work	**Yuhmu (Otomi)**	**Pronunciation classification**	**Acoustic embeddings + SVM/MLP**	**Accuracy, Precision, Recall, and F1-score: 83-94% classification performance**

nonlinear activation function) and Support Vector Machines (SVM, a supervised learning model that seeks to find the optimal hyperplane that maximally separates classes). A grid search (a hyperparameter optimization technique used to find the best combination of parameters in a machine learning model) was applied to identify the ideal hyperparameters for classification.

The analysis was divided into two sections. For the 512 dimensional embeddings, a balanced distribution between male and female speakers was generated, as shown in Table 2, to ensure a controlled dataset for training and testing phases. 66 % of the data was used for training and 33 % for testing, maintaining equitable representation of both genders in each subset. For the 768 dimensional embeddings, a cross-validation approach was employed, which was also part of the hyperparameter search process.

5 Results

This section presents the results obtained during the hyperparameter search for the MLP and SVM models. The optimization process included the evaluation of different values for key parameters, such as the learning rate, number of hidden layers, and neurons in the case of the MLP, as well as the kernel type and regularization parameter for the SVM.

The best results obtained for each model are reported, defined based on performance metrics such as accuracy, precision, recall, and F1-score. These results use vector representations generated by embeddings of sizes 512 and 768.

Table 4 shows the most consistent models identified in the classification of correct and incorrect pronunciation, selected based on their balance among the evaluation metrics obtained through the grid search procedure. Table **??** presents the configurations used in each case, detailing the parameters of the SVM and MLP models employed in the gender-controlled analyses (CG) and in the 6-fold cross-validation (CV).

6 Analysis of Results

This work is distinguished by focusing on the classification of correct and incorrect pronunciation in Yuhmu, a largely unexplored area within the context of MIL (Minority Indigenous Languages), where automatic translation and textual analysis predominate. Compared to previous approaches prioritizing tasks such as ASR (Automatic Speech Recognition) or MT (Machine Translation), this study represents a methodological advance by applying acoustic embeddings to model phonetic patterns.

The results obtained (see Table 3) in both models surpass in consistency those reported in prior work on LRL (Low-Resource Languages) for phonological classification or low-resource ASR tasks.

Unlike generalist models applied to African or Indic languages [8, 26], this study focuses on a highly endangered MIL where acoustic

Table 5. Model configuration summary. CG: Analysis controlled by gender. CV: 6-fold cross-validation using 768-dimensional embeddings. For MLP models, hidden layer configurations indicate the number of layers and neurons per layer

Model	Analysis	Kernel / Activation	Hidden Layers	Reg. / C
Learning & Optimizer				
SVM-1	CG	Polynomial (degree 3)	–	C = 1, γ = scale
–				
MLP-1	CG	ReLU	(30, 80, 80)	α = 0.001 (L2)
Adaptive, Adam				
SVM-2	CV	Linear	–	C = 1
–				
MLP-2	CV	Sigmoid	(55, 55)	α = 0.0001 (L2)
Constant, Adam				

analysis is key for phonological documentation. Likewise, the use of multilingual models such as wav2vec 2.0 [20, 12] and systems like XLSR [6] has shown that knowledge transfer from majority languages improves speech recognition. In this sense, the models implemented in this research partially replicate this logic by utilizing deep speech representations, but applied to a different task: phonetic quality detection.

Furthermore, in the context of Mexico's Indigenous languages, works such as [28, 5, 9] have achieved significant advances in machine translation; however, pronunciation classification remains underdeveloped. Table 5 presents a comparison of previous studies conducted on low-resource and Indigenous languages of Mexico, where our analysis outperforms these works in terms of pronunciation classification accuracy. Therefore, this study contributes to a new research direction that is essential for educational applications such as pronunciation feedback and phonetic training in multilingual contexts, where resources are scarce and languages lack adequate representation in speech technologies.

7 Conclusions and Future Work

The preservation and revitalization of low-resource languages have become increasingly critical in the face of global linguistic homogenization. These languages often lack sufficient digital resources for computational analysis, which hinders their documentation, study, and transmission. Research efforts focused on developing automatic processing and analysis methods for these languages are essential, as they provide the foundation for creating technological solutions that support both learners and native speakers.

The analysis presented in this document is especially relevant in the Mexican context, where, according to [24], only a fraction of MIL have been computationally explored. In this scenario, the use of labeled audio recordings to detect pronunciation errors represents a significant contribution not only to NLP but also to language revitalization efforts, as it enables the development of phonetic support tools that can benefit both speakers and researchers.

In particular, models such as WavLM are considered key tools for obtaining acoustic representations that capture detailed phonetic information, even in low-resource settings such as the one addressed in this work. The base WavLM model, which provides a tensor of embeddings, has shown improved classification performance when the technique of averaging the column values of the tensor is applied. This approach yields more representative and stable embeddings for the task of pronunciation error detection. In this classification task, all available instances are used, and it has been observed that gender-based separation tends to yield lower performance, possibly due to the specificity and variability inherent in such analyses.

Specifically, the phonetic analysis carried out for the Yuhmu language lays the groundwork to

address other linguistic processing aspects in this language, establishing a starting point for further research lines that remain largely unexplored, such as the analysis of specific phonetic conditions affecting pronunciation. Currently, these aspects have not been systematically addressed in computational contexts.

As future work, we propose exploring tasks such as direct audio-to-audio translation without intermediate text, automatic phonetic segmentation, and detailed phoneme-level pronunciation analysis. These research directions would not only expand the scope of developed technologies but also provide key tools for the documentation and preservation of endangered languages.

References

1. **Besacier, L., Barnard, E., Karpov, A., Schultz, T. (2014).** Automatic speech recognition for under-resourced languages: A survey. Speech communication, Vol. 56, pp. 85–100.
2. **Chen, C.-C., Chen, W., Zevallos, R., Ortega, J. E. (2023).** Evaluating self-supervised speech representations for indigenous american languages. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).
3. **Chen, H., Perozzi, B., Al-Rfou, R., Skiena, S. (2018).** A tutorial on network embeddings. arXiv preprint arXiv:1808.02590.
4. **Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. (2022).** Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, Vol. 16, No. 6, pp. 1505–1518.
5. **Chen, W.-r., Abdul-mageed, M. (2023).** Improving neural machine translation of indigenous languages with multilingual transfer learning. **Ojha, A. K., Liu, C.-h., Vylomova, E., Pirinen, F., Abbott, J., Washington, J., Oco, N., Malykh, V., Logacheva, V., Zhao, X.**, editors, Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023), Association for Computational Linguistics, Dubrovnik, Croatia, pp. 73–85. DOI: 10.18653/v1/2023.loresmt-1.6.
6. **Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M. (2021).** Unsupervised cross-lingual representation learning for speech recognition. Interspeech 2021, pp. 2426–2430. DOI: 10.21437/Interspeech.2021-329.
7. **Delprato, M., Frola, A., Antequera, G. (2022).** Indigenous and non-indigenous proficiency gaps for out-of-school and in-school populations: a machine learning approach. International Journal of Educational Development, Vol. 93, pp. 102656.
8. **Doumbouya, M., Einstein, L., Piech, C. (2021).** Using radio archives for low-resource speech recognition: towards an intelligent virtual assistant for illiterate users. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 17, pp. 14757–14765.
9. **Ebrahimi, A., Mager, M., Rijhwani, S., et al. (2023).** Findings of the americasnlp 2023 shared task on machine translation into indigenous languages. Proceedings of the Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP).
10. **Esperanza, M. I. Y. G., Alarcón Montero, R. (2024).** Manual para la escritura de los sonidos del Yuhmu. Secretaría de Cultura, Instituto Nacional de Antropología e Historia, 1 edition.
11. **Europa Press (2022).** Los idiomas, en cifras: ¿cuántas lenguas hay en el mundo?
12. **Kumar, L. A., Dineshraja, V., Naveena, K. S., Renuka, D. K., Resmi, S., Phaniraj, H., Abdul Jabbar, F. (2023).** Self-supervised language identification asr models for low resource indic languages. 2023 International Conference on Modeling, Simulation & Intelligent Computing (MoSICom), IEEE.
13. **Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo,**

G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., Kann, K. (2021). Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Association for Computational Linguistics, Online, pp. 202–217. DOI: 10.18653/v1/2021.americasnlp-1.23.

14. **Magueresse, A., Carles, V., Heetderks, E. (2020).** Low-resource languages: A review of past work and future challenges. arXiv preprint arXiv:2006.07264.

15. **Microsoft (2022).** Wavlm-base-plus-sd. https://huggingface.co/microsoft/wavlm-base-plus-sd. Accessed: 2025-05-17.

16. **Parida, S., Panda, S., Dash, A., Villatoro-Tello, E., Doğruöz, A. S., Ortega-Mendoza, R. M., Hernández, A., Sharma, Y., Motlicek, P. (2021).** Open machine translation for low resource south american languages (americasnlp 2021 shared task contribution). Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Association for Computational Linguistics, Online, pp. 218–223. DOI: 10.18653/v1/2021.americasnlp-1.24.

17. **Penner, K. L. (2019).** Prosodic structure in Ixtayutla Mixtec: Evidence for the foot. Phd thesis, University of Alberta.

18. **Petersen, L., Minkkinen, P., Esbensen, K. H. (2005).** Representative sampling for reliable data analysis: theory of sampling. Chemometrics and intelligent laboratory systems, Vol. 77, No. 1-2, pp. 261–277.

19. **Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X. (2020).** Pre-trained models for natural language processing: A survey. Science China technological sciences, Vol. 63, No. 10, pp. 1872–1897.

20. **Schneider, S., Baevski, A., Collobert, R., Auli, M. (2019).** wav2vec: Unsupervised pre-training for speech recognition. Proc. Interspeech.

21. **Shi, J., Amith, J. D., Chang, X., Dalmia, S., Yan, B., Watanabe, S. (2021).** Highland puebla nahuatl speech translation corpus for endangered language documentation. Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Association for Computational Linguistics, Online, pp. 53–63. DOI: 10.18653/v1/2021.americasnlp-1.7.

22. **Szyszka, M. (2017).** Pronunciation learning strategies and language anxiety. Switzerland: Springer, Vol. 10, pp. 978–3.

23. **Tachbelie, M. Y., Abate, S. T., Schultz, T. (2020).** Development of multilingual asr using globalphone for less-resourced languages: The case of ethiopian languages. Proc. Interspeech, pp. 1032–1036.

24. **Tonja, A., Balouchzahi, F., Butt, S., Kolesnikova, O., Ceballos, H., Gelbukh, A., Solorio, T. (2024).** NLP progress in indigenous Latin American languages. **Duh, K., Gomez, H., Bethard, S.**, editors, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, pp. 6972–6987. DOI: 10.18653/v1/2024.naacl-long.385.

25. **Warjri, S., Pakray, P., Lyngdoh, S. A., Maji, A. K. (2022).** Identification of pos tags for the khasi language based on brill's transformation rule-based tagger. Computación y Sistemas, Vol. 26, No. 2, pp. 989–1005. DOI: 10.13053/cys-26-2-4058.

26. **Woldemariam, Y. (2020).** Transfer learning for less-resourced semitic languages speech recognition: the case of amharic. Proceedings of the 1st joint workshop on spoken language technologies for under-resourced languages (SLTU) and collaboration and computing for under-resourced languages (CCURL), pp. 61–69.

27. Zheng, F., Marrese-Taylor, E., Matsuo, Y. (2024). Improving low-resource machine translation for formosan languages using bilingual lexical resources. Findings of the Association for Computational Linguistics: ACL 2024.

28. Zheng, F., Reid, M., Marrese-Taylor, E., Matsuo, Y. (2021). Low-resource machine translation using cross-lingual language model pretraining. Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas.

Article received on 30/05/2025; accepted on 31/08/2025.
**Corresponding author is J. Arturo Olvera-Lóopez.*