ISSN 1870-4069

A General Overview of Language Pronunciation Analysis Based on Machine Learning

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

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

[2] Laboratorio Nacional de Supercómputo del Sureste de México, Mexico

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

Abstract. Currently, pronunciation analysis is an area related to Natural Language Processing (NLP) which is based on machine learning methods in languages considered universal such as French, English, Mandarin, and Spanish; for low-resource languages such as indigenous languages, some machine learning techniques used to evaluate pronunciation are transfer learning, deep learning, or classic machine learning. Some methods have been applied for pronunciation evaluation, obtaining different levels of performance. This paper provides a review of different approaches, describing phases, languages, metrics, and other important features for the Indigenous Languages from Mexico.

Keywords: Low resource languages, audio analysis, machine learning.

1 Introduction

The goal of Natural Language Processing (NLP) is for computers to understand, interpret, and manipulate human language [26]; an important key is the analysis of pronunciation, which is the correct way in which a word or a language is spoken. This topic is of particular interest to researchers since pronunciation involves articulatory and auditory phonetics, which are sounds of a language that describe its physical aspects and the auditory part analyzes the qualities of sound and describes how it is perceived by the listener [30].

Pronunciation is linked to two skills: speaking and listening in a teaching-learning process; speaking considers the practical and phonological features of a target language (TL), on the other hand, listening refers to the interpretation of the TLs phonological features, which are segments (phonemes) and suprasegments (stress, rhythm and intonation) [30]. The latter are a great challenge for researchers when evaluating isolated words, creating a dilemma

between evaluating pronunciation or global intelligibility, where they diagnose pronunciation and provide feedback on the analyzed words [32].

Computer-Assisted Pronunciation Training (CAPT) applications use different evaluation metrics that help to analyze pronunciation with speaker level scoring methods calculated at the through expressions, words, or local phonemes. Due machine learning pronunciation analysis models and methods require different amounts of data, the analysis of different languages can be categorized into two language groups:

- Universal languages: These are languages that currently have a largest number of speakers worldwide, such as French, English, Mandarin, and Spanish, considering enough instances to be evaluated via machine learning [32].
- Low-resource languages: These are languages with less data, a unique writing system, limited web presence, little understanding of its linguistics, and minimal transcribed speech data and translation dictionaries [4].

In this work, a general overview of the methods used for the analyzing the pronunciation of different languages is presented, describing some corpus, features and classifiers.

2 Pronunciation Analysis

Data gathered on world languages is often unbalanced, considering its number of speakers the data collected on it, so the study and analysis of the NLP has been developed considering two language groups languages. The following section presents a review of the methods and elements used for universal and low-resource languages analysis.

2.1 Universal Languages

Currently, universal languages pronunciation analysis considers data with a high number of instances, in the literature, corpora of this kind contain more than 8,000 instances, with lexical and phonetic descriptors, audio recordings from people of different ages and genders, with acoustic and phonetic knowledge. These data are preprocessed and structured for use in automatic learning. Different corpora have been found in the literature for evaluation tasks as a reference for pronunciation, such as:

- TIMIT which is a corpus that brings together a series of broadband recordings of the English language of 630 speakers from the eight main dialects of American English and which is designed to be implemented in Atomatic Speech Recognition (ASR).
- VoxForge is a corpus composed of 6 languages (English, Spanish, French, German, Russian and Italian) with training, validation, and testing divisions and constant data updating.

Table 1. Some datasets used in related works for audio language analysis.

Paper	Language	Corpus name
Zou, et. al 2018	Mandarin	DidiCallcenter, DidiReading
Duan, et. al 2019	English	Wall Street Journal, LibriSpeech
Feng, et. al 2019	Indo-European	VoxForge, Lwazi
Wang, et. al 2019	Cantonese	CUChild127
Chakroun, et. al 2020	English	TIMIT
Arias-Vergara, et al 2021	German	Verbmobil
Mao, et al 2022	English	Speechocean762, LibriSpeech
Sancinetti, et. al 2022	English	EpaDB
Lin, et. al 2022	English	TIMIT, L2-ARCTIC

Other corpora used in pronunciation analysis are presented in Table 1. These datasets were created for promoting acoustic-phonetic knowledge and automatic speech recognition systems, using phonetic and lexical transcriptions of people of different ages and gender according to their language.

The aforementioned corpora are used to extract relevant features in NLP, in pronunciation, most of the reviewed analyzes use the Mel Frequency Cepstral Coefficients (MFCC), which are a scale that depends on the auditory scale and the coefficients depend on of perception, to be able to duplicate human ears [24] where spectrograms with speech characteristics are obtained, through the Fourier transform, energy power and filter bank that are processed by the discrete cosine transform obtaining cepstral coefficients.

The coefficients can consider different cepstral aspects, as in [7] where 13 MFCCs are extracted with first and second order derivatives representing speech velocity and acceleration, respectively, taking into account 39 cepstral coefficients and integrating Best Tree Encoding (BTE) which is a Wavelet Packet Decomposition (WPD) for ASR and a Image Normalize Encoder (INE).

On the other hand, 14 MFCC coefficients are used in [19] together with their first and second delta, energy (representation of amplitude variations), RMS (the square of the function that defines the continuous waveform), pitch (speed at which the vocal cords vibrate, when pressurized air from the lungs passes through the vocal cords), entropy (a measure of the signal's Fourier power spectrum concentration), spectral features (formants, the most widely used spectral feature, are commonly used to disambiguate vowels and consonants), zero crossing (time domain function which indicates how many times a signal has changed sign with respect to zero) and statistical features.

In [35], 40 MFCC filterbank features, 33-dimensional phonemic posterior features, and 71-dimensional composite posterior features are used for comparison with the deep learning system. The use of Log Phone Posterior (LPP) and Log Posterior Ratio (LPR) has been employed as a vector of phonetic features for the comparative evaluation of phonemes with a high rate of posterior probability per phoneme, proving it to be on par with other methods such as MFCC and with the Word Error Rate (WER) [21]. This feature

vector is used in [8] to train a transformer based on Goodness of Pronunciation features (GOPT) with multitask learning.

The methods used to classify the obtained pronunciation features were initially developed with CAPT based on Hidden Markov Models (HMM) which provide log-likelihood scores, log-posterior with high correlation of human scores and qualifying the pronunciation of a given phoneme and the segment duration score [14]. Some authors implement HMM to detect mispronunciation with the help of ASR based methods, as in [38] in which a model triphone (sequence of three phonemes) was ised to train a Gaussian Mixture Model (GMM) with HMM from left to right with three states.

Some methods for performing pronunciation analysis and feature classification are based on neural networks, one being Deep Neural Networks (DNN), a feedforward artificial neural network with more than one layer of hidden data units between its inputs and outputs. Each hidden unit generally uses the logistic function to map its total input from the lower layer to the scalar state, which it sends to the upper layer [10].

DNN is based on acoustic models, so the network is trained to identify acoustic senones linked to the triphone state. It uses a mixed selection method designed for acoustic modeling based on the GMM, selecting a subset of the senone in the DNN output layer to calculate the posterior probabilities. The senone selection strategy is obtained by grouping the acoustic inputs according to their linear outputs in the hidden upper layer [17].

In [5], DNNs with acoustic-phonetic models are used to detect non-native speech recognition and pronunciation errors and to diagnose articulation-level pronunciation errors of based on the GOP score by calculating the log-posterior ratio between the target canonical phoneme and its most competing phoneme, which has the highest posterior probability for Japanese English learners. Assessment metrics such as False Alarm Rate (FAR), Receiver Operating Characteristic (ROC curve), and Diagnostic Error Rate (DER) are used in this paper, whit 7.82% being the best line error reduction result.

A study carried out in [13] uses a DNN-HMM to improve ASR performance with the Punjabi language, obtaining a Word Error Rate (WER) of 5.32% as the best result. SoftMax uses a DNN to for automatic pronunciation error detection based on GOP; this network is used to carry out transfer learning in order to detect pronunciation errors in Mandarin language using acoustic models trained by different criteria, such as accuracy, F1-score, and Recall[12].

Another pronunciation evaluation method is Convolutional Neural Networks (CNN), which are deep learning architectures inspired by the natural visual perception mechanism of living creatures, where the convolutional layer aims to learn feature representations from inputs; for this type of analysis, this neural network considers digital audio data by feeding the convolution layers with time-frequency representations (spectrograms) of the signals that provide information about how the energy distributed in the frequency domain changes over time [9].

In [3], a CNN is trained to classify speech segments of people with cochlear implants (CI) and healthy control (HC) speakers in the German language, using a cross-validation of k=10 to train and evaluate the models; the performance is measured by means of Presicion (PR), Recovery and F1 score, with the best results obteined from spectrograms of three channels extracted from the compensated transitions, F1 = 0.84. Another study uses a CNN to detect the mispronunciation of phonemes, using forced aligners orthographic transcriptions aligned with the audio recordings automatically generating segmentation at the phoneme level; CNN uses an input channel, output channel, and kernel size of 64, 64, and 9; this analysis uses PR, Recall, and F1-score metrics for evaluating, with F-score showing the best performance with 63.04% [16].

The work proposed in [39] performs an alphabetical classification of the Arabic language, increasing the training data for a better performance of the neural network, adding 20 samples for each alphabet. CNN, Recurrent Neural Network (RNN), and Bidirectional Long Short-Term Memory (BLSTM) are also used; these learning models are accurate to within 91% to 98.5% using Support Vector Machine (SVM) classification.

RNN contains at least one feedback connection, so the activations can flow in a loop, this allows the networks to perform temporal processing and learn sequences [20]; the RNNs are included in CNN to perform a detection and diagnosis of a mispronunciation of the English language; two blocks of CNN and four of RNN with bidirectional LSTM are used, augmentintg the data with acoustic, phonetic, and linguistic embedding (APL) for increased performance; exceeding the baseline by 9.93%, 10.13%, and 6.17% in detection accuracy, DER, and F-measure, respectively [37].

2.2 Low-resource Languages

This section describes the methods used in low-resource languages pronunciation analysis. As mentioned above, this type of language considers a small amount of data, such as text and audio, do not contain enough computational pre-processing instances to be evaluated on their own in a machine learning environment, in some cases as many as a thousand per language.

The use of evaluated corpora in universal languages has been identified in the reviewed literature, with which the neural networks are trained and the subsequent evaluation is carried out with low-resource languages. For example, the LibriSpeech corpus is used in [27] to create ASR in the Tamil language from India and part of Siri Lanka. Another case is the analysis to identify African languages such as Afrikaans, isiNdebele, isiXhosa, isiZulu, Sepedi, Sesotho, Setswana, Siswati, Tshivenda, and Xitsonga, where the VoxForge corpus is used to train a neural network [6].

Another corpus used is PHOIBLE, which is a phoneme database for more than 2,000 languages and dialects, used to develop a tool called Allosaurus that recognizes the phonemes of some universal languages and has been put into practice for low-resource languages [15]. Another language with corpus

preprocessing is Uzbek which uses Common Voice Corpus 8.0 where recordings of sentences on Uzbek dialects are stored [22].

Due to the lack of data on low-resource languages, some authors (in addition to using corpora previously designed for computer environments) have opted to create databases in order to analyze these languages, as in [29], which reports a Tamil and Malay corpus containing 7,582 utterances. Another corpus built to implement a Tibetan speech recognizer contains 28,000 utterances by native speakers [36]. In [23] to recognize the Pashto dialect, a corpus of 900 audio expressions by 45 people was created.

Features similar to those extracted in universal languages can be obtained from audio recordings, such as MFCC [27, 22, 23, 15]. Other features are extracted to analyze low-resource languages with the aim of finding robust features such as the frame-level perceptual linear predictive (PLP) coefficients [6], which are representations conforming to a smoothed short-term spectrum that has been equalized and compressed in a manner similar to that of human hearing [11]. Other features are extracted by [36], where through acoustic features of the speech signal (frequency, amplitude and volume) are obtained through spectrograms.

The features are categorized in different ways when analyzing low-resource languages. One method is transfer learning, applied in [6], where a CNN is trained from a corpus containing 7 Indo-European languages; the weights of the last fully connected layers of the neural network architectures are adjusted using approximately 22 hours of the Lwazi corpus containing 11 African languages, yielding Equal Error Rate (EER) evaluations of 20%; this is a 10% improvement in the identification of African languages (comparison to other authors).

Another method, used for the Uzbek language and its dialects, is End-To-End (E2E) Deep Neural Network-Hidden Markov Model implemented in ASR; it computes the probability from the entire alphabet using the coefficients of MFCC, a deep voice method with CNN and RNN, removing pre-segmented data, and training E2E with Connectionist Temporal Classification (CTC) considering an input voice stream to an output token stream using a single network; this system is capable of training pronunciation, acoustics, and language simultaneously; this system is evaluated with WER, obtaining 16.4% and 17.6% on the Uzbek_Test and Hidden_Test sets, respectively [22].

In [27], a pronunciation analysis is carried out to recognize the Tamil language, where the dialects of the stop words are extracted through the Mel Scales, audio and voice signal features. The features are converted into vectors with LSTM/RNN model and clustered with CNN model to learn interestingness and detect outliers as noise from the original features. Patterns are stored to compare Character Error Rate (CER) and WER through the CNN model.

Another speech approach, applied to the Tibetan language, is that proposed in [36], where one encoder is based on deep CNN and another on a hybrid network of deep CNN and LSTM, it includes a 10-layer CNN architecture. This method uses acoustic features from spectrograms as inputs, computing a WER rating of 36.85%.

Typical classifiers for language recognition are HMM which model a sequence of events or hidden states, Support Vector Machines (SVM) that try to maximize the functional margin between the closest training data from a different class and build an optimal hyperplane, K-Nearest Neighbor (K-NN) which classifies new cases based on a measure of similarity. These classifiers were used for the Pashto dialect which for training and testing purposes are divided into 77% (35 speakers) and 23% (10 speakers), using these classifiers an accuracy of 88%, 84%, and 76% is obtained respectively [23].

Previously developed methods consider more than one language for evaluation and training, generating multilingual recognition methods based on phonetic annotation, such as Allosaurus (Allophone system of automatic recognition for universal speech). This method first calculates the distribution of phonemes using a standard ASR encoder; then, the allophone layer maps the phoneme distribution for each language. This model is trained from start to finish using standard phonemic transcriptions and a list of allophones created by phonetics. The allophone layer is first initialized with the allophone list and then further optimized during the training process.

Allosaurus selected 11 languages (English, Switchboard, Japanese, Mandarin, Tagalog, Turkish, Vietnamese, German, Spanish, Amharic, Italian, Russian) for training with more than 8,000 utterances in each corpus, with 5% of them for testing, the rest were used for validation and training. On the other hand, 2 African languages (Inukitut and Tusom) were used with one thousand randomly selected utterances each; a bidirectional LSTM encoder is used for these, and the phonemes for the training languages are assigned using the grapheme-phoneme tool, creating allophone assignments by specialists in phonetics.

When carrying out an evaluation with the Phoneme Error Rate (PER), accuracy is achieved for the Inuktitut language 84.1% and 77.3% for Tusom. While combined with other corpus (PHOIBLE), error rates are further improved to 73.1% and 64.2% respectively [15].

Indigenous Languages in Mexico. This subsection describes the methods used for the low-resource languages from Mexico, whose analysis has not used machine learning techniques as for African and Asian languages, so the process is still basic compared to the previous ones.

There are sixty-eight indigenous languages in Mexico, distributed throughout national territory, located mainly country's southern and central regions, with linguistic variants producing unique languages in every region. These languages are classified into the following eleven linguistic families: Algica, Yuto-nahua, Cochimi-yumana, Seri, Oto-mangue, Maya, Totonaco-tepehua, Tarasca, Mixe-zoque, Chontal, and Huave [1].

The analysis of indigenous languages from Mexico and other low-resource languages, with limited audio and text data requires different considerations, one of these being transfer learning, which can evaluate them with the support of other data. Another process is to use phonetic and acoustic embeddings from other languages or to perform neural network training using similar phonemes.

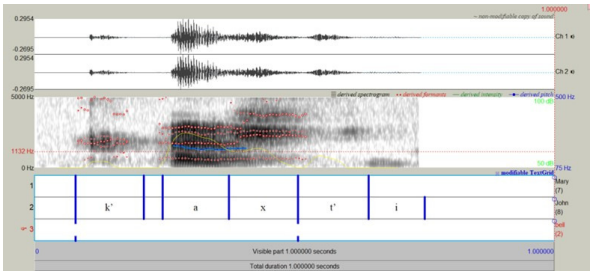

Fig. 1. Representation of analysis with the Praat software of the word yellow said in Otomi.

To carry out pronunciation evaluation processes on indigenous languages of Mexico, various researchers, mostly linguists, have made recordings of words or sentences of people from some communities. These recordings generate corpus of digital audio, with the number of people ranging from five to approximately thirty speakers, as in [25], which considers a corpus of six native Mixtec speakers; on the other hand, [34] uses a database of 8 people with audio recordings of interrogative and declarative sentences in the Otomi language from the Tultepec region in Queretaro; in [28] a corpus of 30 speakers of the Nahuatl language from the state of Puebla is used to carry out an analysis of the production of sonorous sounds; another study carried out by [33] takes into account audio recordings of words uttered by a single person.

Although different language corpora have been mentioned, the methods used for their analysis are similar. A common approach for processing provides as input the audio recordings (without pre-processing or feature extraction) to a software tool such as Praat which is opensource for recording and analyzing words or sentences, computing spectrogram, tone, intensity, volume, and cochleagram(Figure 1).

Another free analytical software is ELAN which represents digital audio in the time from audio or video recordings, it is used to segmentation and made textual annotations to identify phonemes or tones in a subjectively (figure 2); conclusions or evaluations are written qualitatively, commenting on references of tones or between word phoneme similarities of the same language or Spanish.

As it can be seen from figures 1 and 2, the annotations are made under the digital audio time representation and can represent phonemes, syllables, or phrases; unlike Praat, ELAN can process video to obtain audio and analyze complete input sentences; Praat, on the other hand, provides a spectrogram for representing formants as well as for inspecting tones, however, these annotations or analysis are carried out in a unitary and subjective manner by the analyst, how has their own analysis criteria.

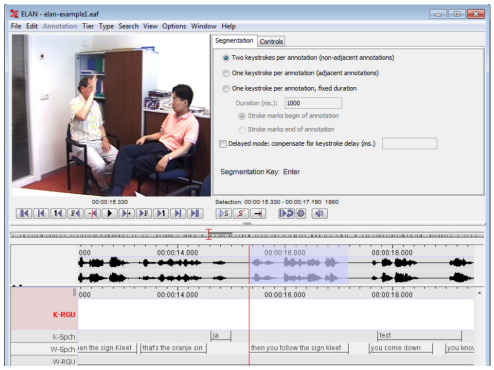

Fig. 2. Analysis representation with ELAN software (Image obtained from [31]).

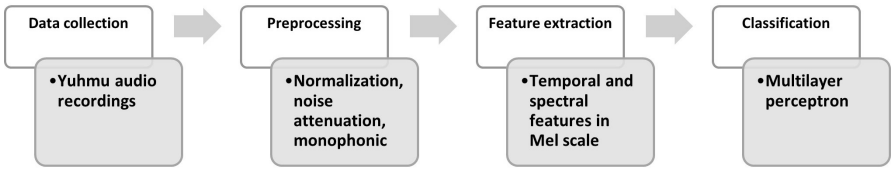

Fig. 3. Phases of the methodology.

3 Proposed Method

The following methodology is proposed for the evaluation of the pronunciation of an indigenous language of Mexico based on the analyzed literature, which considers four important phases (figure 3).

Digital audio recordings of native and non-native people of the Yuhmu indigenous language (a variant of the Otomi language of the State of Tlaxcala) are used in the data collection where Yumhu speakers of 330 words with 3 repetitions each of good pronunciation from [2] is considered, complementing with a corpus of the same words with poor pronunciation of creation own, carrying out a subsampling guided by clustering, removing words with relevant phonetic characteristics (number of phonemes and repetition of phonemes per word), obtaining a final sample of 622 words per category, which consider all the phonemes used in the Yuhmu language.

Within the preprocessing stage, digital audio enhancement is performed, with noise attenuation, amplifying the audio signal and using mono channel for analysis.

During the feature extraction, the use of algorithms based on spectrograms is proposed with experimentation of different parameters for the STFT (Short-time Fourier transform), from which the following are obtained: energy in bands and

Table 2. Audio classification results considering Time, Spectral and Time-Spectral features.

Features	Accuracy (%)
Time	90-91.8
Spectral	91-94.7
Time-Spectral	90-94% and 96-97.7%

shape features; time characteristics (RMS and ZCR), statistics (average and standard deviation) and Pitch.

Finally, a classification of good and bad pronunciation is carried out using a multilayer perceptron considering a "Grid Search" for the search for ideal hyperparameters (one hidden layer, 11 neurons, ReLU activation function, Momentum at 0.4, learning rate at 0.11, and a cross validation with k=5) that would yield the best classification results.

From the previous methodology, 21 characteristic sets are obtained for 4 types of window and overlap with 12 characteristics per set.

The windows used to carry out the experimentation were Hanning, Hamming, Gaussian, and Blackman-Harris, considered for the secondary lobes they have and that can help in the loss of information during the windowing. The window size range used is from 15 to 45 ms with steps of 5 ms, with an overlap of 25, 50, and 75%.

Three analyzes were carried out with different groups of characteristics, the first one considering temporal characteristics, the next one with spectral characteristics and finally using all the characteristics (table 2), to classify good and bad pronunciation.

The results shown in the table 2 present an evaluation of the accuracy of the classification of good and bad pronunciation where the temporal and spectral characteristics show a result of 90-94.7%, while all the characteristics consider two ranges, one similar to the previous ones and another of 96-97.7%, the first is the result of the 15 and 45 ms windows, while when performing a 20-40 ms window the best results are obtained, taking into account that stated in the literature [18]: the best window is within the latter mentioned.

4 Discussion

There are different machine learning processes for evaluating pronunciation due to each language having unique phonetic features, as well as variants in phoneme pronunciation (phono). Figure 4 depicts the methods explained in section 2, showing the researches and the methods used, considering the following:

- Method: Name of the pronunciation evaluation method used.
- Reference: Author(s) of the method.
- Assessment: Metrics used to evaluate the method (WER, PER, PR, EER, Acurracy, DER, CER, among other).

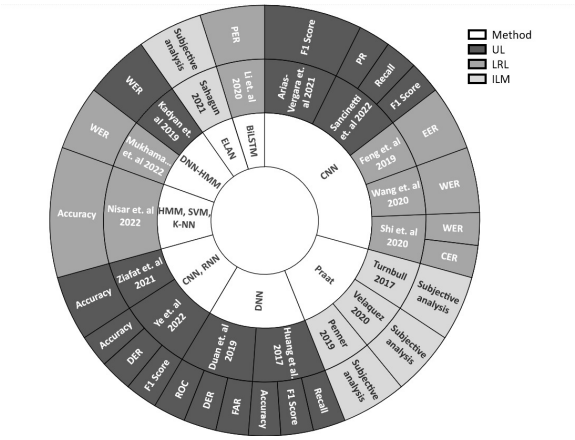

Fig. 4. Methods described in section 2.

– Group: The type of language being analyzed, Universal Languages (UL), Low-resource Languages (LRL), or Indigenous Languages in Mexico (ILM).

From the figure 4, it is evident that the use of neural networks to analyze the UL and LRL is common across all methods, using recurrent, convolutional, and deep neural networks, although in some cases two types of networks are used. The use of machine learning is common; however, as noted above, training is different for of these networks since low-resource languages do not have corpora with more than 8,000 instances as universal languages do. These are aided by a UL corpus for training and a model neural network proposal for classification.

The literature presents some works that apply machine learning for the analysis of indigenous languages for text-to-speech translation tasks, vowel/consonant recognition, however, there are no works that aim to evaluate the pronunciation of the different indigenous languages that exist in Mexico, which presents an area of opportunity in the development of approaches for low-resource languages of Mexico, specifically in pronunciation tasks, because this is still manually generated using software such as Praat or ELAN to perform audio segmentation or interpretation by an analyst, in this case, they are regularly linguists who try to study languages at a semantic level.

From the area referring to assessment, the level of error or accuracy that occurs when making different pronunciations is analyzed, in some cases as in [13, 22, 29, 36] are focused on carrying out an evaluation at the word level considering WER as its measurement, having a better result in the LRL evaluation of 16.4 %. Another metric considered to evaluate is the accuracy that becomes relevant within the methods analyzed, having results even of 98.5% when performing an analysis of universal languages [39].

It is notorious that when observing the type of measurement that is carried out in low-resource languages, a machine learning method to evaluate

pronunciation is not yet considered, this is because currently these processes for this type of language are still under development. Consequently, they are not yet processed like universal languages, due to the number of instances they have, which in some cases reaches a maximum of a thousand, for which reason they have relied on robust corpora for their analysis.

There is a difference between universal languages and low-resource languages, due to the fact that when considering a corpus of a greater number of instances, they are capable of providing a greater number of references for evaluation, which is why the difference in accuracy where the range for world languages is 80-99%, for a machine learning analysis; on the other hand, low-resource languages have results of 60-70% accuracy in their evaluations, having a difference of up to 40 percentage points with respect to universal languages; while the indigenous languages of Mexico have not currently carried out a precision analysis in their processes, so all the results have been concluded with qualitative descriptions.

5 Conclusions

Pronunciation analysis is still a problem of interest for researchers, due to the areas of opportunity that still exist to assess the intelligibility or pronunciation of words or sentences. This machine learning task currently considers two languages groups universal and of low resource, which have been described in this paper.

Different authors have proposed methods with CNN, DNN, RNN or transfer learning, using GOP, EER, PER or WER to define if a word or set of these are well pronounced as evaluation method. On the other hand, the feature extraction methods are similar, considering MFCC, spectrograms, frequencies, energy features, time, among others, of audio recordings. The analysis is carried out with segmentation or supra-segmentation of words or sentences that depend on the type of evaluation, considering phonemes, articulation zone or tone of the same.

It is worth noting that low-resource languages analysis does not consider a method of analysis through machine learning similar to that of universal languages where they can be evaluated autonomously, while the indigenous languages of Mexico are considered a challenge for researchers when evaluating pronunciation.

When carrying out the experimentation with the Yuhmu language, it is concluded that the characteristics used for other types of languages are considered useful to analyze this type of languages. Currently, the analysis of the pronunciation of the indigenous languages of Mexico would help the conservation, preservation and dignification of this kind of languages.

Acknowledgments. This work was supported by the National Council of Humanities Science and Technology (CONAHCYT) under the scholarship number 814401 and the 189 VIEP-BUAP project.

References

1. Catálogo de las lenguas indígenas nacionales (2022), https://www.inali.gob.mx/clin-inali, last accessed May 06, 2023
2. Alarcon Montero, R.: Manual para la escritura de los sonidos del yuhmu. INAH (2023)
3. Arias-Vergara, T., Klumpp, P., Vasquez-Correa, J. C., Nöth, E., Orozco-Arroyave, J. R., Schuster, M.: Multi-channel spectrograms for speech processing applications using deep learning methods. Pattern Analysis and Applications, vol. 24, pp. 423–431 (2021)
4. Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech communication, vol. 56, pp. 85–100 (2014)
5. Duan, R., Kawahara, T., Dantsuji, M., Nanjo, H.: Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 391–401 (2019)
6. Feng, K., Chaspari, T.: Low-resource language identification from speech using transfer learning. In: 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2019)
7. Gbaily, M. O.: Automatic database segmentation using hybrid spectrum-visual approach. The Egyptian Journal of Language Engineering, vol. 8, no. 2, pp. 28–43 (2021)
8. Gong, Y., Chen, Z., Chu, I.-H., Chang, P., Glass, J.: Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7262–7266. IEEE (2022)
9. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al.: Recent advances in convolutional neural networks. Pattern recognition, vol. 77, pp. 354–377 (2018)
10. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, vol. 29, no. 6, pp. 82–97 (2012)
11. Hönig, F., Stemmer, G., Hacker, C., Brugnara, F.: Revising perceptual linear prediction (plp). In: Ninth European Conference on Speech Communication and Technology (2005)
12. Huang, H., Xu, H., Hu, Y., Zhou, G.: A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection. The Journal of the Acoustical Society of America, vol. 142, no. 5, pp. 3165–3177 (2017)
13. Kadyan, V., Mantri, A., Aggarwal, R., Singh, A.: A comparative study of deep neural network based punjabi-asr system. International Journal of Speech Technology, vol. 22, pp. 111–119 (2019)
14. Kim, Y., Franco, H., Neumeyer, L.: Automatic pronunciation scoring of specific phone segments for language instruction. In: Fifth European Conference on Speech Communication and Technology (1997)
15. Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., Anastasopoulos, A., Mortensen, D. R., Neubig, G., Black, A. W., et al.: 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. IEEE (2020)

16. Lin, B., Wang, L.: Phoneme mispronunciation detection by jointly learning to align. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6822–6826. IEEE (2022)
17. Liu, J.-H., Ling, Z.-H., Wei, S., Hu, G.-P., Dai, L.-R.: Cluster-based senone selection for the efficient calculation of deep neural network acoustic models. In: 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). pp. 1–5. IEEE (2016)
18. Liu, L., He, J., Palm, G.: Effects of phase on the perception of intervocalic stop consonants. speech communication, vol. 22, no. 4, pp. 403–417 (1997)
19. Maqsood, M., Habib, H. A., Nawaz, T.: An efficientmis pronunciation detection system using discriminative acoustic phonetic features for arabic consonants. Int. Arab J. Inf. Technol., vol. 16, no. 2, pp. 242–250 (2019)
20. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: Interspeech. vol. 2, pp. 1045–1048. Makuhari (2010)
21. Minh, N. Q., Hung, P. D.: The system for detecting vietnamese mispronunciation. In: Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications: 8th International Conference, FDSE 2021, Virtual Event, November 24–26, 2021, Proceedings 8. pp. 452–459. Springer (2021)
22. Mukhamadiyev, A., Khujayarov, I., Djuraev, O., Cho, J.: Automatic speech recognition method based on deep learning approaches for uzbek language. Sensors, vol. 22, no. 10, pp. 3683 (2022)
23. Nisar, S., Tariq, M.: Dialect recognition for low resource language using an adaptive filter bank. International Journal of Wavelets, Multiresolution and Information Processing, vol. 16, no. 04, pp. 1850031 (2018)
24. 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 (2020)
25. Penner, K.: Prosodic structure in ixtayutla mixtec: Evidence for the foot, (2019)
26. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: A survey. Science China Technological Sciences, vol. 63, no. 10, pp. 1872–1897 (2020)
27. Rajendran, S., Mathivanan, S. K., Jayagopal, P., Venkatasen, M., Pandi, T., Sorakaya Somanathan, M., Thangaval, M., Mani, P.: Language dialect based speech emotion recognition through deep learning techniques. International Journal of Speech Technology, vol. 24, pp. 625–635 (2021)
28. Sahagun, A. S.: Spanish VOT Production by L1 Nahuat Speakers. Ph.D. thesis, University of Saskatchewan (2021)
29. Shi, K., Tan, K. M., Duan, R., Salleh, S. U. M., Suhaimi, N. F. A., Vellu, R., Thai, N. T. H. H., Chen, N. F.: Computer-assisted language learning system: Automatic speech evaluation for children learning malay and tamil. In: INTERSPEECH. pp. 1019–1020 (2020)
30. Szyszka, M.: Pronunciation learning strategies and language anxiety. Switzerland: Springer, vol. 10, pp. 978–3 (2017)
31. Tacchetti, M.: User's guide for elan linguistic annotator. The Language Archive, MPI for Psycholinguistics, Nijmegen, The Netherlands.[Google Scholar], (2017)
32. Tejedor-García, C., Escudero-Mancebo, D., Cámara-Arenas, E., González-Ferreras, C., Cardeñoso-Payo, V.: Assessing pronunciation improvement in students of english using a controlled computer-assisted pronunciation tool. IEEE Transactions on Learning Technologies, vol. 13, no. 2, pp. 269–282 (2020)

33. Turnbull, R.: The phonetics and phonology of lexical prosody in san jerónimo acazulco otomi. Journal of the International Phonetic Association, vol. 47, no. 3, pp. 251–282 (2017)
34. 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, vol. 8, no. 2, pp. 143–168 (2020)
35. Wang, J., Qin, Y., Peng, Z., Lee, T.: Child speech disorder detection with siamese recurrent network using speech attribute features. In: INTERSPEECH. vol. 2, pp. 3885–3889 (2019)
36. Wang, W., Yang, X., Yang, H.: End-to-end low-resource speech recognition with a deep cnn-lstm encoder. In: 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP). pp. 158–162. IEEE (2020)
37. Ye, W., Mao, S., Soong, F., Wu, W., Xia, Y., Tien, J., Wu, Z.: An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (apl) embeddings. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6827–6831. IEEE (2022)
38. Zhang, Z., Wang, Y., Yang, J.: Masked acoustic unit for mispronunciation detection and correction. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6832–6836. IEEE (2022)
39. Ziafat, N., Ahmad, H. F., Fatima, I., Zia, M., Alhumam, A., Rajpoot, K.: Correct pronunciation detection of the arabic alphabet using deep learning. Applied Sciences, vol. 11, no. 6, pp. 2508 (2021)