Goodness of Pronunciation Algorithm in the Speech Analysis and Assessment for Detecting Errors in Acoustic Phonetics: An exploratory review

Edward Wilder Caro Anzola*, *Member, IEEE*, and Miguel Ángel Mendoza Moreno**, *Member, IEEE*

Abstract*— The correct word pronunciation in an oral communication language is important to reach a total message comprehension that is transmitted; all languages in the world have certain rules that a native or non-native speaker must comply in a conversational process; however, some people have deficiencies when they are learning a foreign language or, in the case of diseases, patients have difficulties in the production of speech sounds. Technological solutions are implemented around the new algorithms' developments, such as the Goodness of Pronunciation (GOP) approach and its variants that enhance the performance and accuracy. GOP uses a probabilistic approximation to calculate the likelihood ratio between canonical and spoken phonemes in assessment tasks, but it is dependent of speech corpus characteristics, signal multidimensionality and thresholding of scores. The actual research is centered on GOP evaluation through the comparison between baseline-GOP and others speech evaluation mechanisms based on Machine Learning (ML) and/or Neural Networks (NN). The scores on metrics have improved with the input data training that increases factors like accuracy, correlation ratio, database resolution, samples generalization, error discriminative ability, among others. The background was collected through a basic exploration of "GOP" and "Goodness of Pronunciation" keywords in co-occurrence analysis mapping of VOSviewer® software; then, a descriptor table was implemented as a binnacle. The contrast among GOP algorithm and other strategies reveals that GOP algorithm is limited to phones and segmental phonemes and there is not suitable for applying in supra-segmental components; however, the spread corpus obtained from data speech treatment in previous stages of GOP calculation can be enhanced the overall performance and accuracy, extended to the prosodic analysis in complex studies. All in all, GOP method can be applied to assess principal components of speech samples, especially, in medical issues to detect mispronunciations for some diseases and impairments. The principal objective of this study is to evaluate the GOP algorithm proficiency to carry out a portable device based on evolutionary algorithm classifier integrated in an embedded system for children mispronunciation diagnosis in phonoaudiology therapy.

***Index Terms*— Algorithm, Automatic Speech Recognition, Goodness of Pronunciation, scoring, speech mispronunciation.**

I. INTRODUCTION

THE pronunciation quality is expressed in the assessments obtained from the articulated, coherent and correct emission of typical sounds of a language; however, the mechanical sound production is the same regardless of the language that is being spoken. The human phonatory system presents characteristics that are the object of study by phonoaudiology experts since two aspects: (i) the analysis of the articulatory, acoustic and auditory phonetic elements from the mechano-sound production and reception and (ii) the phonological analysis from the expressive functionality for a suitable communication.

In the development of linguistics and neurolinguistics, as applied sciences, two primordial concepts, that are not very common, are defined: *orthology*, that establishes conventional norms for the correct pronunciation of phonemes, and *orthophony*, that defines methods for correcting and improving the pronunciation [1][2][3]; both aspects are important for the evaluation and treatment of several speech characteristics, such as the acoustic production of vowels, consonants and their variants in segmental phonetics, and intonation, rhythm, duration, frequency, intensity and stress at suprasegmental phonetics level, this last aspect is relevant in the phonology area [4][5][6].

Pronunciation errors affect human communication from the unintelligibility of the messages that are emitted; these problems may be due to a bad oral practice (acquired) or to a physiological and/or morphological impairments; for this reason, the professional in speech therapy determines the quality of pronunciation through the evaluation of speech tests where a diagnosis is obtained, it allows some corrective actions through an appropriate treatment [7][8]. An alternative to study the quality of pronunciation is the Goodness of Pronunciation algorithm (GOP) developed by Silke Maren Witt in her doctoral thesis dissertation; in this work, she proposed to obtain a scoring on the pronunciation between phones samples and a phonetic corpus transcribed by phoneticians; these tests yielded scores that consider the pronunciation as correct or incorrect through a statistical

The authors contributed equally to this work.

*Edward Wilder Caro Anzola is with Pedagogical and Technological University of Colombia, Avenida Central del Norte 39-115, Tunja, Boyacá, Colombia (e-mail: edward.caro@uptc.edu.co). ORCID: 0000-0002-6602-3665

**Miguel Ángel Mendoza Moreno is with Pedagogical and Technological University of Colombia, Avenida Central del Norte 39-115, Tunja, Boyacá, Colombia (e-mail: miguel.mendoza@uptc.edu.co). ORCID: 0000-0001-9000-5881

evaluation of the phonemes and their correlations [9].

In this work, we propose a background review that is centered on the GOP algorithm, its scope and improvement, the relationship with Automatic Speech Recognition systems (ASR) and the correlative discussion from authors' findings. The main objective of this document is the compilation of the conceptual foundations required for the doctoral thesis proposal emphasis on the error detection in the Colombian Spanish phonetic pronunciation in a child population through evolutionary algorithms supported by artificial intelligence (AI) embedded systems; the proposal will be developed between TELEMATICs and I^{2}E research groups, which belong to the Faculty of Engineering of the Pedagogical and Technological University of Colombia.

II. The Basic GOP Algorithm

Measure pronunciation quality is a relevant process for non-native speakers who want to learn another language; technology provides the ability to learn pronunciation from tools based on computer systems, such as Computer-Assisted Language Learning (CALL) or Computer-Assisted Pronunciation Training (CAPT) resources [10][11]. The GOP algorithm presents an effective way for measuring the quality of pronunciation from the statistical comparison between captured phones or phones records and a speech corpus endorsed by specialists in phonetics through the assess of systematic deviation obtained (weighted errors); Witt and Young [12], establish a relationship between the acoustic observed segments (named *O(q)*) and phones samples (named *q*) for a correspondence of likelihoods; being *NF(p)* the frames of the acoustic segments, *p* the acquired phones and *Q* a group of phone models obtained by Hidden Markov Model processing (HMM), the likelihood is expressed by:

$$\boldsymbol{GOP_1(p)} = \left| log\left(\frac{p(O^{(p)}|p)P(p)}{\sum_{q\in Q} p(O^{(p)}|q)P(q)}\right)\right| \Big/ NF(p). \qquad (1)$$

If this relationship takes on a good likelihood approximation between *P(p)* and *P(q)*, numerator and denominator will be calculated through an alignment forced between sequenced transcriptions of phone models and iterative operations of unrestricted phones, respectively. Alignment forced algorithm proposed by Viterbi [13], is an important process to achieve the phone boundaries of sequences in corpus database and an adequate segmentation of acquired speech; thus, a comparison between the structures can be achieved. According to Witt and Young, there are some criterions to bear in mind in the GOP algorithm application: (i) when an error is presented, log likelihood in denominator must ensure that the frame under inspection does not have constrains, this guarantees a maximum likelihood estimation, (ii) formant constructions differ between native and non-native speakers, (iii) human judgments are subjective in speech studies, (iv) training database needs variable thresholds for the extraction of speech characteristics, (v) mispronunciations can be categorized by phoneme/word substitution or by unknowing pronunciation, and (vi) GOP performance can be evaluated using four statements: strictness in error identification, agreement between reference and automatic transcriptions, cross-correlation between marked reference and automatic detected errors, and overall statistical reject correlation between phone errors.

GOP algorithm, as a pronunciation scoring paradigm, has a good potential to find phonemic and prosodic errors in speech samples determined by the deletion/insertion of phonemes or by the constitutive accent components, respectively; for this case, each feature pronunciation component makes part of a representative multidimensional space; therefore, it can be measured and scored [14][15]. In the beginning, GOP was implemented as a tool to determine the mispronunciations in foreign language learners; however, its application has been extended to another fields like medicine, where it is very useful to detect speech diagnosis anomalies, therapy and rehabilitation progress or degenerative diseases evolution; Table I collects some categories of GOP´s usage in our background research (obtained from VOSviewer®).

TABLE I
Fields Of Application for the GOP Framework In Issues Related To Mispronunciations Detection

Application fields	Objective of GOP	References
Foreign language learning or Pronunciation proficiency	Mandarin/Chinese mispronunciation detection	[17], [19], [20], [21], [26], [30], [31], [44], [54], [57], [59], [66], [67], [69]
	Dutch non-native speakers error detection	[18], [24], [25], [35]
	Technology for language and education	[22], [29], [42], [76], [86], [89], [98]
	English teaching and learning	[16], [23], [27], [28], [34], [40], [56], [64], [65], [70]
	Other language mispronunciation issue (Arabic, Spanish, Vietnamese)	[38], [49], [85]
	Improvement of GOP algorithm	[32], [41], [45], [46], [53], [55], [62], [68], [71], [72], [73], [79], [81], [82], [84], [87], [90], [91], [93], [95], [96], [99]
	Comparison between GOP and other algorithms	[33], [36], [37], [39], [48], [50], [51], [52], [60], [74], [75], [78], [88], [92], [94]
Medicine: diagnosis, therapy and/or rehabilitation	Morphologic or physiologic pathology/injury	[43], [77], [80]
	Speech comprehensibility	[47], [74]
	Improvement of GOP process for therapy	[58], [97]
	Childhood apraxia	[61], [63]
	GOP combined with another medical signals	[83]

III. GOP as a Technological Tool in Language Education

Actually, technology is the mediator between learning objectives and real outcomes in educational processes, especially, in language learning; hardware and software converge in practical solutions around proficiency and improvement in the knowledge acquisition on foreign language. ASR computational tool is the principal scaffold to build development environments for speech analysis and synthesis procedures, such as CAPT or CALL ecosystems; Spoken Language Technology for Education (SLATE) appears as a requirement to work in all forms of spoken

language excluding corporal or sign language [22]. CALL and CAPT platforms provide the possibility to practice and receive feedback to compare and correct the pronunciation in different architectures since speech recognition and error detection [29][98]. GOP is useful in CAPT platform due to the capacity to calculate scores in recognition and free phone loops stages [42]. Pronunciation and intonation evaluation were evaluated through GOP and specialized software like PRAAT due to big data volume with enough information that allows to extract other characteristics like fluency, stress and rhythm [76]. Another software that offers efficiency in phonetic analysis is Kaldi, an open-source toolkit, that combined with the characteristics of GOP, allows to build enhanced forced alignment and more accurate metrics [89].

IV. GOP Algorithm Improvements

The GOP algorithm is not infallible, the errors in score computing depend of multiple factors like environment of sound capture, instrumental offsets, algorithm precision and constraints, hardware and software requirements, subjectivity of human evaluation, among others. Sometimes, the GOP method is enough with the teacher's experience without the necessity of an expert database, a model can be obtained with teachers' linguist knowledge to create an error pattern database [31] or by marginal statistical distribution of raw speech signals [71]. The noise is one of the problems in acoustic signal capture during the speech evaluation due to its energy or frequency components can overlap the speech information, for this reason, noise reduction techniques are important in speech pre-processing; in analog signals, n-order filters are implemented with several active circuits that reduce overall energy sound in thresholds levels, whereas, in digital signals, software resources are implemented since FIR-IIR filters until noise reduction algorithms (NR), for instance, the Stereo-based Piecewise Linear Compensation for Environments method (SPLICE) [32]. GOP is used to characterize the segmental level of non-native speech, but at suprasegmental level, the measures of fluency and articulation rate of speech, phonation ratio or length of pauses need other algorithms like subspace Gaussian Mixture Models (GMM) [41]; extension approaches in GOP, such as multi-view, multi-granularity [95], and multi-aspect [91] allow to evaluate prosodic through self-supervised learning (SSL) features [87]. In conventional GOP, based on GMM-HMM, lattice search space and threshold scores limit the GOP performance, because of this, the use of genetic algorithms like Deep Neural Networks (DNN), improve the GOP calculations through refinement of the correlation of high dimensional features, increase of non-linear mapping, training of model parameters and the generalization capability [45]. The acoustic models obtained by GOP-GMM-HMM method need a convenient model space for training data, thus, the insertion of new criterion like Maximum F1-Score Discriminative Training can be an improve to refine the Word Error Rate (WER) [46] and the overall accuracy [55].

One of the shortcomings detected in GOP is the direct comparison between the "ideal" and the pronounced phoneme, but it does not include combined scores for prosodic, fluency, completeness or accuracy components; however, the use of context-aware GOP (CaGOP) [72] and the Dynamic Time Warping algorithm (DTW) allow to analyze non-synchronous time sequences and pitch frequencies like a multidimensional hyperparameters [53][96]. The original GOP is dependent of the phoneme context in the basic GMM-HMM and others GOP improvement algorithms of the statistical inference like estimation, classification or error where state transition probabilities and sub-phonemics acoustic models (senones) are calculated separately; a conjunction of HMM transitions with DNN-HMM score computations increase the GOP correlation capability [62]. GOP, as part of CAPT in non-native speech, needs human knowledge information and big corpora database to build an acoustic model, for this reason, a transfer learning method is a good strategy to provide acoustic models based on phonetic and articulatory features [73][79]. The phonemes require segmental boundaries that are needed by the canonical and pronounced phones alignment in segmentation procedures, but this alignment is not precise due the non-proportional relationship between numerator and denominator in GOP equation, one solution is proposed by the use of Bidirectional Long Short-Term Memory (BDLSTM) method that performs the spatial and sequential encoding of short acoustic segments [81]; another proposal is to integrate the speaker metadata of the phonetician to the BDLSTM resource which allows to expand the information of phonetic corpus [86]. One solution for the numerator/denominator discriminative relationship is the Maximum Mutual Information (MMI) estimation that maximizes the reference transcriptions probability while decrease the probability of other alternatives [82]. The sentence pronunciation assessment is another problem for GOP because it is not possible to obtain a good score when phonemes are chained in a phrase, GOP is designed for a segmental analysis, but not for suprasegmental cases, then GOP variants are proposed, the average-GOP (aGOP) and the confusion-GOP (cGOP) statistical computation approaches, the results present increasing performances compared to conventional GOP [84].

The GOP algorithm may not be suitable for extracting all speech features due to the limited corpus database, utterance variations (stress, phonation, stress) of voice samples taken as canonical, extraction features algorithms, among others; the inclusion of DNN-HMM optimized with lattice-free maximum mutual information (LFMMI) criterion and a feature learning module perform the GOP robustness [90][93]. The use of native pronunciation like a reference for data speech analysis is a requirement to compare the non-native pronunciation in GOP; pre-training models in GOP previous stages allow to enhance the performance in regard to baseline-GOP model through a convolutional layer including in the deep feedforward sequential memory network and HMM algorithm (DFSMN-HMM) to determine the linguistic-acoustic similarity [99]. The GOP augmentation capability is a good strategy in pronunciation error detection with the fine-grained phoneme-level augmentation approach, named as SpeechBlender, where a multi-task learning

is added to extract transitions and styles in speech information that are masking and creating new dataset space of mispronunciation and accent variants [94].

V. Comparison between GOP and other Algorithms

GOP algorithm works with frame datasets of speech samples declared as canonical pronunciation database (referenced phonetics in a corpus), then, a statistical comparison generates a score based on data deviation; however, baseline-GOP suffers errors in the approximation because its dependence of input data. Previously, we described some research that sought to improve the performance and robustness of GOP using HMM/GMM mixtures with neural networks or optimization algorithms to increase the data precision through data training models and/or refined features extraction methods. GOP calculates the overall similarity between phonemes through a discriminative strategy, but at phone and word levels, nonconformities are presented, so a new two-pass discriminative assessment (TPDA) is proposed to improve the measure of the scoring difference and the Pearson's correlation; the comparative results show more matches in TPDA method [33]. The speech accent is an undetectable parameter in GOP being an insignificant feature in the final scoring; therefore, an automatic accent algorithm is needed, this is achieved through Weighted Finite State Transducer (WFST) pronunciation resource in an alignment phase and Maximum Entropy (ME) estimation technique in a scoring phase, then GOP is surpassed in this shortcoming [36]. The application of neural networks and learning algorithms enhance the detection and classification of the phoneme compared with GOP capabilities; some tests with Logistic Regression (LR) classifier overcomes the recall and precision rate with a relative improvement respect to GOP [37].

DTW was compared with GOP, verifying the minimum distance between segment location in time, two different versions of DTW are proposed; first, DTW error detection to find similar segments and, second, DTW thresholding to determine the minimal distance and normalize the segment [39]; later, four approaches are defined, Variations of the Word Structure (VoWS), Normalized Phoneme Distances Thresholding (NPDT), Furthest Segment Search (FSS) and Normalized Furthest Segment Search (NFSS) where the mispronounced detection was more fast with a higher accuracy for limited corpus [48]. Voice and aspiration sound characteristics are included in another study where segmental and sub-segmental errors are wanted through phone-based and attribute-based classifiers, the Equal Error Rate (EER), Detection Accuracy (DEA), Diagnostic Accuracy (DA), precision and recall are slightly better than GOP scoring [50]. If the canonical and pronounced phonemes are defined as probability distributions, they have a relative entropy error that can be calculated through an empathized deviation expressed by the Kullback-Leibler Divergence (KLD) measure that performs the senone vectors like stochastic speech segments in a phonetics data space constructed by DNN methods; however, GOP has not an important difference score compared with the metric proposal [51]. Although, baseline-GOP is based on HMM-GMM approach, four variants of weighted GMM algorithms are realized for comparation, in addition, Phonetic Distance Classifier (PDC) are included in the comparative process; not all algorithms performed well, only one aims to overcome at GOP due to tuned thresholds added [52].

GOP can be potentialized when is combined with training networks based on DNN or Convolutional Neural Networks (CNN); a special approach, named Siamese Neural Network (SNN) is structured around three metrics: Diagnostic Accuracy (DA), False Rejection Rate (FRR) and False Acceptance Rate (FAR) to determine the mispronunciation at phone level using high level embedding representations; SNN has a better accuracy with variable thresholds [60] (results complemented by the same authors in [75]). GOP has another disadvantage, the rhythm and pauses of speech samples are not detectable and the spectrum can have leakage, so, the use of windowing filter like Hamming window and with the design of a cognitive heuristic computational algorithm, allow to obtain signal spectrum smoothing and extend the capability of phoneme/word classification and the accuracy of error detection [92]. The use of deep feature clustering through the classification of phonemes using K-nearest neighbor (KNN), Naïve Bayes (NB) and SVM approaches report a high accuracy detection, besides, the use of unsupervised Phone Variation Model (PVM) guarantees the implementation of prosodic error detections that GOP does not have [78]. The proposed algorithms are not always efficient because they are totally dependent of size and quality of corpus database, also, for the same language, the regional accents can have noticeable differences that confuse the segmentation or extraction features algorithms; ASR-based Pronunciation Error Detection (PED) is contrastable with variants of GOP method, the results have high variability when speech dataset is based on one accent, but WER metric increases when an accent mixtures in the phonetical corpus are added [88].

VI. Applications of GOP Method

A. GOP in Foreign Language Learning

In the language teaching/learning environments, the improvement of speaking skill is important to develop an effective communication with native people from other countries; however, language components like accent, intonation or conditionate sound generation of vocal tract can be a problem for a good learning process in beginners or non-native speakers. Classical methods of language learning require an environment with the direct participation of a teacher (that is the "knowledge expert") and a repetitive training about vocabulary with correct phonetic transcriptions from the student, it represents time consuming and big investment; for this reason, computational systems based on ASR used by CALL or CAPT are presented as good practical solutions [16]. The quality of pronunciation in foreign language learners is relevant for a fluently and concise communication, especially, in languages with high tonal contents (e.g. Chinese Mandarin) that need special feedback since the assessment protocols, in this way, Goodness of Pronunciation becomes Goodness of Tone, an alternative to find phonetic and prosodic errors through frequency variation feature vectors

represented in multiple space distribution and operated through HMM or GMM [17][44].

Chinese is one of the languages that has used GOP algorithms to assess mispronunciations in non-native speakers; syllables and phones are scoring using the nature of GOP, it is, log-likelihood results that are obtained as a calculating deviation between speaker and canonical pronunciation at principal frequency formant F0 level and other components; corpus data can be obtained through own experimental protocols or using external speech databases; however, good algorithm performances are reported [19][20]. GOP has been used as a baseline test on Chinese mispronunciation detection, if tone errors are not including, other applied methods in own databases are reliable; these cases are shown in researches that support findings on different feature extraction algorithms and resources like pronunciation models on Support Vector Machines (PSM-SVM) [21], fast aligned GOP algorithm as basic GOP modification (FAGOP) [26], structural features on SVM [30], combination between Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) estimations [54], extended speech recognition with neural networks (DNN-ERN) [57], audio signal, face video and 3-D animation correlation [59], bidirectional Long Short-Term Memory network associated with GOP and tone labels [66], integration between DNN and HMM in a Maximum Evaluation Criterion Training (MECT) [67], separation of initial and final syllables in GOP process [69]; all experimental results had contrasted with basic GOP performance and a mixture of GOP with neural networks or HMM-GMM procedures had reported better behaviors about mispronunciation errors detection and segmental speech classification (more sensitivity and resolution).

English language is a necessity for global communication in many society fields (academics, business, politics, social networks, etc.); according to the British Council report "The Future of English: Global Perspectives" (https://futureofenglish.britishcouncil.org/). Nowadays, English as a Second Language (ESL) is the dominant requirement for people that need to interact effectively with the world; with almost 2 billion of total speakers, 83% are foreign language learners (L2). Automatic test in proficiency English language is a complementary tool for teachers or human raters that searches a good discriminatory level in utterance assessment for determining accuracy, rate of speech (ROS) and GOP scores [23]. The length of the sentences in a machine scoring can produce nonconformities in a basic GOP calculation, for this reason, bounded lattices with more tokens in lattice based GOP (LGOP) are proposed for improving the sentences correlation [27]; even more, if an English student has a strong accentuation in his/her native language (e.g. Indian speakers), the speech dataset used for training models will influence the phone and prosodic attributes, that can decrease the GOP accuracy [34]; another case, is viewed in English learning by Korean students where, there are many variants for certainly phonemes compared with canonical utterances, so, a mixture between logistic regression and GOP methods are applied to enhance the prediction/scoring errors in a Generalized Transformation-Based error-driven Learning (GTBL) approach [40]. Chinese students consider English like their second language, but the pronunciation is affected by marked accents, so, some changes in the GOP algorithm based on modified Maximum Likelihood Linear Regression (MLLR) were applied for determining an adaptative data in the acoustic model adjust given more accuracy in segmental error detection [56]; likewise, automatic speech resources are proposed to apply in English speaking tests in schools through DNN-HMM acoustic adaptation technique combined with GOP [64] and teacher utterance based GOP (TGOP) [68], both in noisy environments; besides, in primary school to track the English word pronunciation, GOP-DNN-HMM algorithms are implemented on Hidden Markov Model Toolkit (HTK) [65], all processes give good scoring features; DNN-HMM integrated in GOP method has allowed to implement dataset corpus through refined pronunciation assessment (RPA) focused on pronounced/canonical phoneme and particular/overall score pronunciation with enhanced accuracy [70]. Japanese learners use automatic speech errors to determine the speaker sounds quality since the variant acoustic models in speech structures [28]; another metric, called reference-free error rate (RER), compared GOP with GMM-HMM-DNN evaluation systems where genetic algorithms obtain better human/machine correlations [38].

Other foreign languages teaching methods use GOP statistical procedure to evaluate non-native mispronunciations, as it is the case of Dutch as a second language (L2) where there are some speech mistakes due substitution of fricative sound by a plosive sound, GOP is compared with other three computational classifiers: Weigelt algorithm, Linear Discriminant Analysis with acoustic phonetic features (LDA-APF) and Linear Discriminant Analysis with Mel Frequency Cepstrum Coefficients (LDA-MFCC); LDA classifiers had slightly higher Scoring Accuracy (SA) than GOP [18][24]; however, with adequate thresholds, GOP increases the performance for Dutch consonant pronunciation [25], and vowels sounds [35]. Finally, GOP is used in speech error detection and performance measures for phonemes and utterance in native Arabic language [38], non-native Spanish Japanese learners [49] and tonal vowels Vietnamese students [85].

B. GOP in Medical Diagnosis and Therapeutics Issues

One of the capabilities that GOP offers is its use in medical diagnosis of speech; some speech impairments are a direct consequence of an articulatory system trauma or by defective development of anatomical components of sounds production, on the other hand, many neuromotor impairments can affect the vocal sounds indirectly and their evolution is an important indicative value for degenerative processes. Facial palsy is a temporal weakness of facial muscles due to partial damages of the nerve that control their movements, for this reason, the adequate phonation of complete phrases is difficult to realize and the communication ability decrease; some labial, labiodental and fricative phonemes are affected, hence, the

GOP method is a good proposal to determine the facial palsy grade according to disease evolution and the communication ability of patients since the motor muscles that control the phonological articulatory system [43]. Another disease that is common, especially in children, is the cleft lip and palate that produce dysphonia (characterized by breathiness, hoarseness, and low intensity of voice); the use of two-dimensional discrete cosine transform (2D-DCT) that feeds a SVM algorithm, allows obtaining the vowel onset points (VOPs) strategy that is compared with GOP, where VOP overcomes GOP in misarticulated stops during the diagnosis speech tasks [77]. The correct phoneme pronunciation totally depends of the position of articulatory organs, some speech problems arise when the air is forced to pass across airway tract incorrectly producing frequency changes, GOP adding with a gammatonegram study allows to find spectral and entropy changes in a spectrogram, so, the therapist can know the errors sounds, for example, phoneme substitutions [74].

Speech impairments due to mild traumatic brain injury (mTBI) are analyzed since the variations of speech patterns in patients with post-traumatic headache (PTH), in this case, the GOP scores are estimated by the precision in a vowels and consonants pronounced in a motivated text, others normalized and average measures are taken; here, gender and age are not discriminative elements, but during a headache episode, rhythm and speech rates decrease the speech capacity [80]. Forced-aligned GOP (F-GOP) is an algorithm variant proposed to evaluate intelligibility and comprehensibility in speech sound disorders (SSD) that can be applied to multiple correlated diseases [47]. At an early age, some children have speech disorders that will affect their communication skills in adulthood, so therapists and phonoaudiologist use technological resources, such as GOP to diagnose the impairment and make treatment decisions; two kinds of GOP are proposed as mispronunciation detector, correct and incorrect patterns GOP (GOP-CI) and applied SVM in GOP (GOP-SVM), both results are contrasted with the performance of baseline LGOP through the Corpus of Children's Pronunciation (CCP), the GOP algorithms investigated have improvements around Unweighted Average Recall (UAR), balanced label recognition rate and high correlation with human experts judgements [58]. In medicine, the acoustic signal brings important information about mechanical sound behavior; however, other electric and non-electric biosignals contain useful data to complete an accurate diagnosis, like electrophysiological markers, diagnostic imaging, chemical analysis, among others; the use of ultrasound imaging in tongue during a speech task is reported as a good mechanism to detect some speech disorders and there are high matches between device results and those reported by the therapist [83].

The speech ability is analyzed from the neuroscience research view; one expression of the human though is the speech communication and the brain must be coded these complex synaptic processes to become their signals in articulated sounds; some studies are centered on finding the relationship between neural biosignals and phonemes to use the information in psychophysiological mechanisms speech analysis, human behavior and psychologic emotions, brain computer interaction (BCI) applications, new requirements to build hardware solutions in speech treatment, neurolinguistics and speech correction, among others; GOP can be complemented with the mechanisms described above [97]. Apraxia is a neurologic disorder that presents difficulty speaking due to uncoordinated movements and loss of sound sequence, this problem is generated by brain injuries that needs an especial treatment; in this disease, the identification of deviations between correct and disorder speech corpus through GOP offers a good verification of phoneme-level pronunciations through DNN classifiers like One-Class SVM model [61][63].

VII. Conclusions

In this review, GOP algorithm was explored since its conception, application, improvement and comparison with a specific purpose: to observe the ability of GOP to determine the quality of speech samples to define errors in pronunciation tasks; GOP is a good approach to find phoneme error at segmental level according to the phonetic transcription tables that are known by experts in the field of speech therapy, their knowledge is important to implement a phonetic corpus, especially, the canonical database. GOP compares probabilities between accepted utterances and spoken words/phrases obtained in signal segmentation and feature extraction processes, but it is limited by corpus size, thresholds in scoring and other components like accent, intonation, noise, alignment between phonemes and other constrains. Baseline-GOP algorithm is effective in phones and syllables; however, for some extended speech samples, the performance is poor; for this reason, the investigations centered on GOP have been improved through the use of intelligent algorithms derived from Machine Learning and/or Neural Networks where the primary dataset is applied to train robust acoustic models that enhanced the space of database features.

GOP applications are widely worked in foreign language learning; however, the background around the medical trends shows the usage extensibility in speech disorders diagnosis and therapy. This review is an initial part of our future project where the GOP algorithm will be evaluated by its implementation in embedded hardware systems like artificial intelligence Single Board Computers (AI-SBC) based on a Tensor Processing Unit (TPU) or a Neural Processing Unit (NPU). Furthermore, a new proposal to integrate GOP score machine with evolutionary algorithms, for example, Learning Classifier System approach (LCS), will be designed in an ubiquitous model for phonoaudiology usage in children speech impairments detection.

Acknowledgment

This work would not have been possible without the support of TELEMATICs and I^2E research groups of Pedagogical and Technological University of Colombia.

REFERENCES

[1] A. M. Salvador Rosa, "En defensa de la Ortología," in *Boletín de la Real Academia Española*, Tomo 95, Cuaderno 311, 2015, pp. 179-190.

[2] E. Roldán, "Sobre la ortología," in *Revista Documentos Lingüísticos y Literarios* UACh, Número 37, 2018, pp. 215-218.

[3] A. Lorenc, "Diagnosis of the pronunciation norm," in *Logopedia,* Issue 42, Research Project No. 2012/05/E/HS2/03770, pp. 61-86.

[4] J. H. Clegg, and W. C. Fails, "Manual de fonética y fonología españolas," in *Routledge Introductions to Spanish Language and Linguistics*, 1st ed., G. Parodi, P. Cantos-Gómez and C. Howe Ed., 2018, ch 8, pp. 123-136.

[5] E. C. Zsiga, "Approaches to the Interface," in *The Phonology/Phonetics Interface*, P. Ackema, M. Ota Ed., Edinburgh Advanced Textbooks in Linguistics, 2020, pp. 7-23.

[6] I. Rumyantseva, "Phonetics and Phonology as the Two Aspects of one Science of Human Speech Sounding," in *Psycholing.*, vol. 23, no. 2, pp. 203–213, Apr. 2018, doi: 10.5281/zenodo.1199220.

[7] A. Ygual-Fernández, and J. F. Cervera-Mérida, "Relación entre la percepción y la articulación en procesos fonológicos sustitutorios de niños con trastornos del lenguaje," *Rev. Neurol. 2013*, vol. 56, no. 1, pp. S131-S140, Feb. 2013, doi: 10.33588/rn.56S01.2013012.

[8] M. H. Franciscatto *et al*., "Towards a speech therapy support system based on phonological processes early detection," in *Computer Speech & Language*, vol. 65, Jan. 2021, pp. 1-20, Art. no. 101130, doi: 10.1016/j.csl.2020.101130.

[9] S. M. Witt, "Use of Speech Recognition in Computer-assisted Language Learning," Ph.D. dissertation, Dep. Eng., Cambridge Univ., Cambridge, England, 1999.

[10] X. Chen, D. Zou, H. Xie, and F. Su, "Twenty-five years of computer-assisted language learning: A topic modeling analysis," in *Language Learning & Technology*, vol. 25, no. 3, pp. 151-185, Oct. 2021. [Online]. Available: http://hdl.handle.net/10125/73454

[11] H. S. Mahdi, and A. A. Al Khateeb, "The effectiveness of computer-assisted pronunciation training: A meta-analysis," in *BERA: Review of Education*, vol. 7, no. 3, Oct. 2019, pp. 733-753, doi: 10.1002/rev3.3165

[12] S. M. Witt, and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," in *Speech Communication*, vol. 30, no. 1-2, pp. 95-108, Feb. 2000, doi: 10.1016/S0167-6393(99)00044-8

[13] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260-269, April 1967, doi: 10.1109/TIT.1967.1054010

[14] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub (1999). "Automatic scoring of pronunciation quality," in *Speech Communication*, vol. 30, no. 2-3, pp. 83-93, Feb. 2000, doi: 10.1016/S0167-6393(99)00046-1

[15] S. M. Witt, "Automatic Error Detection in Pronunciation Training: Where we are and where we need to go," in *Proc. ISADEPT*, Stockholm, Sweden, 2012, pp. 1-8

[16] S. V. Lohiya, and M. V. Kamble, "Survey on Computer Aided Language Learning using Automatic Accent Assessment Techniques," *2015 International Conference on Pervasive Computing (ICPC)*, Pune, India, 2015, pp. 1-4, doi: 10.1109/PERVASIVE.2015.7087089

[17] L. Zhang, *et al.*, "Automatic Detection of Tone Mispronunciation in Mandarin," in *Proc. ISCSLP*, Singapore, Singapore, 2006, pp. 590-601, doi: 10.1007/11939993_61

[18] H. Strik, K. Truong, F. Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," presented at the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), Antwerp, Belgium, Aug 27-31, 2007, doi: 10.21437/Interspeech.2007-512

[19] J. Zheng, C. Huang, M. Chu, F. K. Soong, and W. Ye, "Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Honolulu, HI, USA, Apr 15-20, 2007, pp. 201-204, doi: 10.1109/ICASSP10710.2007

[20] C. Huang, F. Zhang, and F. K. Soong, "Improving Automatic Evaluation of Mandarin Pronunciation with Speaker Adaptive Training (SAT) and MLLR Speaker Adaption," *2008 6th International Symposium on Chinese Spoken Language Processing - ISCSLP 2008*, Kunming, China, Dec 16-19, 2008, pp. 1-4, doi: 10.1109/iscslp14491.2008

[21] S. Wei, G. Hu, Y. Hu, and R. H. Wang, "A new Method for Mispronunciation Detection using Support Vector Machine based on Pronunciation Space Models," in *Speech Communication*, vol. 51, no. 10, pp. 896-905, October 2009, doi: 10.1016/j.specom.2009.03.004

[22] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832-844, October 2009, doi: 10.1016/j.specom.2009.04.005

[23] F. Wet, C. Van der Walt, and T. R. Niesler, "Automatic assessment of oral language proficiency and listening comprehension," *Speech Communication*, vol. 51, no. 10, pp. 864-874, October 2009, doi: 10.1016/j.specom.2009.03.002

[24] H. Strik, K. Truong, F. Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845-852, October 2009, doi: 10.1016/j.specom.2009.05.007

[25] S. Kanters, C. Cucchiarini, and H. Strik, "The Goodness of Pronunciation Algorithm: A Detailed Performance Study," in *Proc. SLaTE*, Warwickshire, England, 2009, pp. 49-52

[26] C. Liu, F. Pan, F. Ge, B. Dong, and Y. Yan, "Forward Optimal Measures for Automatic Mispronunciation Detection," in *Proc. ISCSLP*, Tainan, Taiwan, 2010, pp. 80-84, doi: 10.1109/ISCSLP16673.2010

[27] Y. Song, W. Liang, and R. Liu, "Lattice-based GOP in automatic pronunciation evaluation," in *Proc. ICCAE*, Singapore, Singapore, 2010, pp. 598-602, doi: 10.1109/ICCAE16361.2010

[28] N. Minematsu, S. Asakawa, M. Suzuki, and Y. Qiao, "Speech Structure and Its Application to Robust Speech Processing," in *New Gener. Comput*., vol. 28, Aug. 2010, pp. 299-319, doi: 10.1007/s00354-009-0091-y

[29] J. Doremalen, H. Strik, and C. Cucchiarini, "Speech Technology in CALL: The Essential Role of Adaptation," in *Proc. ITEC*, Kortrijk, Belgium, 2010, pp. 56-69, doi: 10.1007/978-3-642-20074-8_5

[30] T. Zhao, A. Hoshino, M. Suzuki, M. Minematsu, and K. Hirose, "Automatic Chinese Pronunciation Error Detection Using SVM Trained with Structural Features," *2012 IEEE Spoken Language Technology Workshop (SLT)*, Miami, FL, USA, Dec 2-5, 2012, pp. 473-478, doi: 10.1109/SLT20433.2012

[31] Y. B. Wang, and L. S. Lee, "Improved Approaches of Modeling and Detecting Error Patterns with Empirical Analysis for Computer-Aided Pronunciation Training," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar 25-30, 2012, pp. 5049-5052, doi: 10.1109/ICASSP.2012.6289055

[32] Y. Luan, M. Suzuki, Y. Yamauchi, N. Minematsu, S. Kato, and K. Hirose, "Performance Improvement of Automatic Pronunciation Assessment in a Noisy Classroom," *2012 IEEE Spoken Language Technology Workshop (SLT)*, Miami, FL, USA, Dec 2-5, 2012, pp. 428-431, doi: 10.1109/SLT.2012.6424262

[33] I. Zhang, F. Pan, B. Dong, and Y. Yan, "A Novel Discriminative Method for Pronunciation Quality Assessment," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 26-31, 2013, pp. 8223-8226, doi: 10.1109/ICASSP.2013.6639268

[34] S. Joshi, and P. Rao, "Acoustic Models for Pronunciation Assessment of Vowels of Indian English," in *Proc. O-COCOSDA/CASLRE*, Gurgaon, India, 2013, pp. 1-6, doi: 10.1109/ICSDA.2013.6709904

[35] J. Doremalen, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch," in *J. Acoust. Soc. Am.*, vol. 134, no. 2, August 2013, pp. 1336–1347, doi: 10.1121/1.4813304

[36] F. William, A. Sangwan, and J. H. L. Hansen, "Automatic Accent Assessment Using Phonetic Mismatch and Human Perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, September 2013, pp. 1818-1829, doi: 10.1109/TASL.2013.2258011

[37] W. Hu, Y. Qian, and F. K. Soong, "A New Neural Network Based Logistic Regression Classifier for Improving Mispronunciation Detection of L2 Language Learners," *The 9th International Symposium on Chinese Spoken Language Processing*, Singapore, Sept 12-14, 2014, pp. 245-249, doi: 10.1109/ISCSLP.2014.6936712

[38] A. A. Hindi, M. Alsulaiman, G. Muhammad, and S. Al-Kahtani, "Automatic Pronunciation Error Detection of Nonnative Arabic Speech," *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, Doha, Qatar, Nov 10-13, pp. 190-197, doi: 10.1109/AICCSA.2014.7073198

[39] M. Bugdol, Z. Segiet, and M. Kręcichwost, "Pronunciation Error Detection Using Dynamic Time Warping Algorithm," in *Information Technologies in Biomedicine*, vol. 4, pp. 345-354, doi: 10.1007/978-3-319-06596-0_32

[40] E. Bang, J. Lee, G. Lee, and M. Chung, "Pronunciation Variants Prediction Method to Detect Mispronunciations by Korean Learners of English," *ACM Transactions on Asian Language Information Processing*, vol. 13, no. 4, pp 1-21, Art. no. 16, doi: 10.1145/2629545

[41] R. Tong, B. P. Lim, N. F. Chen, B. Ma, and H. Li, "Subspace Gaussian Mixture Model for Computer-Assisted Language Learning," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 4-9, 2014, pp. 5347-5351, doi: 10.1109/ICASSP.2014.6854624

[42] I. Odriozola, L. Serrano, I. Hernaez, and E. Navas, "The AhoSR Automatic Speech Recognition System," in *Proc. IberSPEECH*, Las Palmas, España, 2014, pp. 279-288, doi: 10.1007/978-3-319-13623-3_29

[43] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, "The Goodness of Pronunciation algorithm applied to disordered speech," presented at the 15th Annual Conference of the International Speech Communication Association (Interspeech 2014), Singapore, Sep 14-18, 2014, doi: 10.21437/Interspeech.2014-357

[44] R. Tong, N. F. Chen, B. Ma, and H. Li, "Goodness of Tone (GOT) for Non-native Mandarin Tone Recognition," presented at the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Dresden, Germany, Sep 6-10, 2015, doi: 10.21437/Interspeech.2015-254

[45] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers," *Speech Communication*, vol. 67, pp. 154-166, March 2015, doi: 10.1016/j.specom.2014.12.008

[46] H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum F1-Score Discriminative Training Criterion for Automatic Mispronunciation Detection". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, April 2015, pp. 787-797, doi: 10.1109/TASLP.2015.2409733

[47]L. Fontan, T. Pellegrini, J. Olcoz, and A. Abad, "Predicting disordered speech comprehensibility from Goodness of Pronunciation scores," presented at the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Dresden, Germany, Sep 6-10, 2015, doi: 10.18653/v1/W15-5108

[48] Z. Miodonska, M. D. Bugdol, M. and Krecichwost, "Dynamic time warping in phoneme modeling for fast pronunciation error detection," in *Computers in Biology and Medicine*, vol. 69, February 2016, pp. 277-285, doi: 10.1016/j.compbiomed.2015.12.004

[49] V. Álvarez, D. Escudero, C. González, and V. Cardeñoso, "Evaluating Different Non-native Pronunciation Scoring Metrics with the Japanese Speakers of the SAMPLE Corpus," in *Proc. IberSPEECH*, Lisboa, Portugal, 2016, pp. 205-214, doi: 10.1007/978-3-319-49169-1_20

[50] W. Li, S. M. Siniscalchi, N. F. Chen, and C. H. Lee, "Improving Non-Native Mispronunciation Detection and Enriching Diagnostic Feedback with DNN-Based Speech Attribute Modeling," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, Mar 20-25, 2016, pp. 6135-6139, doi: 10.1109/ICASSP.2016.7472856

[51] W. Hu, and F. K. Soong, "KL-Divergence based Mispronunciation Detection via DNN and Decision Tree in the Phonetic Space," in *Proc. APSIPA*, Jeju, Korea (South), Dec 13-16, 2016, doi: 10.1109/APSIPA.2016.7820849

[52] S. Robertson, C. Munteanu, and G. Penn, "Pronunciation Error Detection for New Language Learners," presented at the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016), San Francisco, CA, USA, Sept 8-12, 2016, doi: 10.21437/Interspeech.2016-539

[53] K. Sheoran, *et al*., "Pronunciation Scoring with Goodness of Pronunciation and Dynamic Time Warping," *IEEE Access*, vol. 11, 2023, pp. 15485-15495, doi: 10.1109/ACCESS.2023.3244393

[54] G. Huang, H. Li, R. Zhou, and Y. Zhou, "A Text-Independent Method for Estimating Pronunciation Quality of Chinese Students," in *Information Technology and Intelligent Transportation Systems*, vol. 69, Xi'an, China, June 10, 2017, pp. 201-211, doi: 10.1007/978-3-319-38771-0_20

[55] H. Huang, H. Xu, Y. Hu, and G. Zhou, "A Transfer Learning Approach to Goodness of Pronunciation based Automatic Mispronunciation Detection," in *J. Acoust. Soc. Am.*, vol. 142, no. 5, November 2017, pp. 3165-3177, doi: 10.1121/1.5011159

[56] G. Huang, *et al*., "English Mispronunciation Detection Based on Improved GOP Methods for Chinese Students," *2017 International Conference on Progress in Informatics and Computing (PIC)*, Nanjing, China, Dec 15-17, 2017, pp. 425-429, doi: 10.1109/PIC.2017.8359585

[57] W. Li, S. M. Siniscalchi, N. F. Chen, and C. H. Lee, "Using Tone-based Extended Recognition Network to Detect Non-native Mandarin Tone Mispronunciations," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, Korea (South), Dec 13-16, 2016, pp. 1-4, doi: 10.1109/APSIPA38779.2016

[58] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," in *Computer Speech & Language*, vol. 50, July 2018, pp. 62-84, doi: 10.1016/j.csl.2017.12.006

[59] X. Peng, H. Chen, L. Wang, and H. Wang, "Evaluating a 3-D virtual talking head on pronunciation learning," in *International Journal of Human-Computer Studies*, vol. 109, January 2018, pp. 26-40, doi: 10.1016/j.ijhcs.2017.08.001

[60] Z. Wang, J. Zhang, and Y. Xie, "L2 Mispronunciation Verification Based on Acoustic Phone Embedding and Siamese Networks". *11th International Symposium on Chinese Spoken Language Processing (ISCSLP 2018)*, Taipei, Taiwan, Nov 26-29, 2018, pp. 444-448, doi: 10.1109/ISCSLP.2018.8706597

[61] M. Shahin, J. Ji, and B. Ahmed, "One-Class SVMs Based Pronunciation Verification Approach," *24th International Conference on Pattern Recognition (ICPR 2018)*, Beijing, China, Aug 20-24, 2018, pp. 2881-2886, doi: 10.1109/ICPR.2018.8545687

[62] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An Improved Goodness of Pronunciation (GoP) Measure for Pronunciation Evaluation with DNN-HMM System Considering HMM Transition Probabilities," *20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep 15-19, 2019, pp. 954-958, doi: 10.21437/Interspeech.2019-2363

[63] M. Shahin, and B. Ahmed, "Anomaly detection based pronunciation verification approach using speech attribute features," in *Speech Communication*, vol. 111, August 2019, pp. 29-43, doi: 10.1016/j.specom.2019.06.003.

[64] D. Luo, L. Xia, C. Zhang, and L. Wang, "Automatic Pronunciation Evaluation in High-states English Speaking Tests Based on Deep Neural Network Models," *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China, 2019, pp. 124-128, doi: 10.1109/ICAIBD.2019.8836976

[65] H. Wan, J. Xu, H. Ge, and Y. Wang, "Design and implementation of an English pronunciation scoring system for pupils based on DNN-HMM," *10th International Conference on Information Technology in Medicine and Education (ITME 2019)*, Qingdao, China, Aug 23-25, 2019, pp. 348-352, doi: 10.1109/ITME.2019.00085.

[66] L. Wei, N. F. Chen, S. M. Siniscalchi, and C. H. Lee, "Improving Mispronunciation Detection of Mandarin Tones for Non-Native Learners with Soft-Target Tone Labels and BLSTM-Based Deep Tone Models," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, Apr 15-20, 2018, pp. 2012-2024, doi: 10.1109/ICASSP34228.2018.

[67] B. Chen, and Y. C. Hsu, "Mandarin Chinese Mispronunciation Detection and Diagnosis Leveraging Deep Neural Network Based Acoustic Modeling and Training Techniques," in *Computational and Corpus Approaches to Chinese Language Learning*, pp. 217-224, February 2019, doi: 10.1007/978-981-13-3570-9_11.

[68] S. Sudhakara, M. K. Ramanathi, C. Yarra, A. Das, and P. K. Ghosh, "Noise Robust Goodness of Pronunciation Measures using Teacher's Utterance," in *Proc. SLaTE*, Graz, Austria, 2019, pp. 69-73, doi: 10.21437/SLaTE.2019-13.

[69] W. Dong, and Y. Xie, "Normalization of GOP for Chinese Mispronunciation Detection," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, Nov 18-21, 2019, pp. 1004-1008, doi: 10.1109/APSIPAASC47483.2019.

[70] Y. Gao, *et al.*, "XDF-REPA: A Densely Labeled Dataset toward Refined Pronunciation Assessment for English Learning," in *Proc. O-COCOSDA*, Cebu, Philippines, Oct 25-27, 2019, pp. 1-6, doi: 10.1109/O-COCOSDA46868.2019.9041154.

[71] S. Cheng, *et al.*, "ASR-Free Pronunciation Assessment," presented at the 21st Annual Conference of the International Speech Communication Association, (Interspeech 2020), Shanghai, China, Oct 25-29, 2020, doi: 10.21437/Interspeech.2020-2623.

[72] J. Shi, N. Huo, and Q. Jin, "Context-aware Goodness of Pronunciation for Computer-Assisted Pronunciation Training," presented at the 21st Annual Conference of the International Speech Communication Association, (Interspeech 2020), Shanghai, China, Oct 25-29, 2020, doi: 10.48550/arXiv.2008.08647.

[73] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "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, doi: 10.1109/TASLP.2019.2955858

[74] P. B. Ramteke, S. Supanekar, V. Aithal, and S. G. Koolagudi, "Identification of Palatal Fricative Fronting using Shannon Entropy of Spectrogram," in *Proc. MIKE*, Goa, India, Dec 19–22, 2019, pp. 234-243, doi: 10.1007/978-3-030-66187-8_22

[75] Y. Xie, Z. Wang, and K. Fu, "L2 Mispronunciation Verification Based on Acoustic Phone Embedding and Siamese Networks," in *J. Sign. Process. Syst.*, September 2020, pp. 1-11, doi: 10.1007/s11265-020-01598-z

[76] J. Dong, Y. Liao, X. Li, and W. Huang, "The Application of Big Data to Improve Pronunciation and Intonation Evaluation in Foreign Language Learning," in *Proc. ICHSA*, Kunming, China, Jul 20-22, 2019, pp. 160-168, doi: 10.1007/978-3-030-31967-0_18

[77] V. Mathad, and S. Mahadeva, "Vowel Onset Point Based Screening of Misarticulated Stops in Cleft Lip and Palate Speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 450-460, doi: 10.1109/TASLP.2019.2957887

[78] F. Nazir, M. Majeed, M. Ghazanfar, and M. Maqsood, "A computer-aided speech analytics approach for pronunciation feedback using deep feature clustering," in *Multimedia Systems*, vol. 29, pp. 1699-1715, July 2021, doi: 10.1007/s00530-021-00822-5

[79] M. Sancinetti, J. Vidal, C. Bonomi, and L. Ferrer, "A Transfer Learning Approach for Pronunciation Scoring," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, Singapore, May 23-27, 2022, pp. 6812-6816, doi: 10.1109/ICASSP43922.2022.9747727

[80] C. Chong, *et al*., "Altered speech patterns in subjects with post-traumatic headache due to mild traumatic brain injury," in *J. Headache Pain*, vol. 22, no. 82, July 2021, pp. 1-12, doi: 10.1186/s10194-021-01296-6

[81] J. A. Lopez Saenz, A. Jalal, R. Milner, and T. Hain, "Attention Based Model for Segmental Pronunciation Error Detection," *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Cartagena, Colombia, Dec 13-17, 2021, pp. 725-732, doi: 10.1109/ASRU51503.2021.9687993

[82] D. Luo, M. Guan, and L. Xia, "Automatic Scoring of L2 English Speech Based on DNN Acoustic Models with Lattice-Free MMI," in *Proc. MLICOM*, Shenzhen, China, Sep 26-27, 2020, pp. 113-122, doi: 10.1007/978-3-030-66785-6_13

[83] M. Ribeiro, J. Cleland, A. Eshky, K. Richmond, S. Renals, "Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors," in *Speech Communication*, vol. 128, April 2021, pp. 24-34, doi: 10.1016/j.specom.2021.02.001

[84] B. Su, *et al*., "Improving Pronunciation Assessment Via Ordinal Regression with Anchored Reference Samples," *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun 6-11, 2021, pp. 7748-7752, doi: 10.1109/ICASSP39728.2021.9413659

[85] N. Quang Minh, and P. Duy Hung, "The System for Detecting Vietnamese Mispronunciation," in *Proc. FDSE*, Singapore, Singapore, Nov 24-26, 2021, pp. 452-459, doi: 10.1007/978-981-16-8062-5_32

[86] J. A. Lopez Saenz, and T. Hain, "Use of Speaker Metadata for Improving Automatic Pronunciation Assessment," in *Proc. SLSP*, Cardiff, UK, Nov 23-25, 2021, pp. 61-72, doi: 10.1007/978-3-030-89579-2_6

[87] F. Chao, T. Lo, T. Wu, Y. Sung, and B. Chen, "3M: An Effective Multi-view, Multi-granularity, and Multi-aspect Modeling Approach to English Pronunciation Assessment," In *Proc. APSIPA ASC*, Chiang Mai, Thailand, Nov 7-10, 2022, pp. 575-582, doi: 10.23919/APSIPAASC55919.2022.9979979

[88] X. Wei, C. Cucchiarini, R. van Hout, and H. Strik, "Automatic Speech Recognition and Pronunciation Error Detection of Dutch Non-native Speech: cumulating speech resources in a pluricentric language," in *Speech Communication*, vol. 144, October 2022, pp. 1-9, doi: 10.1016/j.specom.2022.08.004

[89] C. Batista, A. Dias, and N. Neto, "Free resources for forced phonetic alignment in Brazilian Portuguese based on Kaldi toolkit," *EURASIP J. Adv. Signal Process.*, February 2022, pp. 1-32, doi: 10.1186/s13634-022-00844-9

[90] B. Lin, and L. Wang, "Gated fusion of handcrafted and deep features for robust automatic pronunciation assessment," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2022)*, Chiang Mai, Thailand, Nov 7-10, 2022, pp. 1399-1404, doi: 10.23919/APSIPAASC55919.2022.9979961

[91] H. Do, Y. Kim, and G. Lee, "Hierarchical Pronunciation Assessment with Multi-Aspect Attention," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, Jun 4-10, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095733

[92] Y. Wu, C. Zheng, M. Hao, and L. Wang, "Implementation of a System for Assessing the Quality of Spoken English Pronunciation Based on Cognitive Heuristic Computing," in *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 5239375, pp. 1-12, doi: 10.1155/2022/5239375

[93] D. Luo, L. Xia, and M. Guan, "Noise Robust Automatic Scoring Based on Deep Neural Network Acoustic Models with Lattice-Free MMI and Factorized Adaptation," in *Mobile Networks and Applications*, Volume 27, pp. 1604-1611, doi: 10.1007/s11036-021-01878-3

[94] Y. El Kheir, S. Chowdhury, H. Mubarak, S. Afzal, and A. Ali, "Speechblender: Speech Augmentation Framework for Mispronunciation Data Generation," Arxiv., to be published, Cornell University, November 2022, pp. 1-5, doi: 10.48550/arXiv.2211.00923

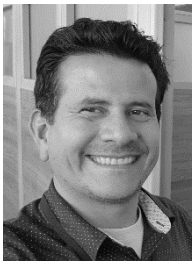
[95] Y. Gong, Z. Chen, I. Chu, P. Chang, and J. Glass, "Transformer-Based Multi-Aspect Multi-Granularity Non-Native English Speaker Pronunciation Assessment," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 23-27, 2022, pp. 7262-7266, doi: 10.1109/ICASSP43922.2022.9746743

[96] J. Duan, and Z. He, "An English Pronunciation and Intonation Evaluation Method based on the DTW Algorithm," in *Soft Computing*, vol. 27, no. 6, March 2023, pp. 1-9, doi: 10.1007/s00500-023-08027-w

[97] A. Morkovina, A. Shevchenko, V. Stroganova, and A. Vartanov, "Analysis of psychophysiological mechanisms and approaches to the correction of pronunciation," in *Natsional'nyy psikhologocheskiy zhurnal (National psychological journal)*, vol. 1, no. 49, 2023, pp. 77–87, doi: 10.11621/npj.2023.0107

[98] E. Cámara-Arenas, C. Tejedor-García, C. Tomas-Vázquez, and D. Escudero-Mancebo, "Automatic pronunciation assessment vs. automatic speech recognition: A study of conflicting conditions for L2-English," in *Language Learning & Technology*, vol. 27, no. 1, pp. 1-19

[99] W. Liu, *et al*., "Leveraging Phone-Level Linguistic-Acoustic Similarity for Utterance-Level Pronunciation Scoring," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, , Jun 4-10, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096699



Edward Wilder Caro Anzola received the B.S. degree in electronic engineering from the Pedagogical and Technological University of Colombia (UPTC), Sogamoso, Colombia, in 2002, the Sp. Degree in electronic instrumentation from Santo Tomás University, Tunja, Colombia, in 2007. He is currently development the Ph.D. in engineering with emphasis in systems and computing engineering from the UPTC, Tunja, Colombia. He is currently a professor and researcher at the Electronic Engineering School, UPTC, Tunja, Colombia. His research interests include embedded systems programming, speech signal processing, electronic bioinstrumentation and evolutionary computation.

Miguel Ángel Mendoza Moreno received the B.S. degree in systems and computer engineering from the Universidad Pedagógica y Tecnológica de Colombia (UPTC), Tunja, Colombia, the M.Sc. degree in science from the Ciencias de la Información y las Comunicaciones, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia, Sp. in Redes y Servicios Telemáticos form the Universidad del Cauca, Colombia, Sp. in Pedagogía para el Desarrollo del Aprendizaje Autónomo from the Universidad Nacional Abierta y a Distancia, Colombia, the Ph.D. degree in Ciencias de la Electrónica from the Universidad del Cauca, Colombia. He is currently a professor and researcher at the Systems Engineering School, UPTC, Tunja, Colombia. The work topics related with his works are include adaptive systems, personalization of learning, Internet of Things, smart cities and telematics in general.