IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Pronunciation Assessment and Automated Analysis of Speech in Individuals with Down Syndrome: Phonetic and Fluency Dimensions

Mario Corrales-Astorgano[1], *César González-Ferreras*[1], *David Escudero-Mancebo*[1], *Lourdes Aguilar*[2], *Valle Flores-Lucas*[1], *Valentín Cardeñoso-Payo*[1], *Carlos Vivaracho-Pascual*[1]

[1]Universidad de Valladolid, Spain [2]Universitat Autònoma de Barcelona, Spain

mcorrales@infor.uva.es, cesargf@uva.es, descuder@infor.uva.es, Lourdes.Aguilar@uab.cat, mariavalle.flores@uva.es, valentin.cardenoso@uva.es, cevp@infor.uva.es

Abstract

In this study, we analyze the potential use of an annotated corpus to identify various dimensions of speech quality, including phonetics and fluency, in individuals with Down syndrome, enabling the development of automated assessment systems. Two experiments were conducted: for phonetic evaluation, we used the Goodness of Pronunciation (GoP) metric with an automatic segmentation system and correlated results with a speech therapist's evaluations, showing a positive trend despite not notably high correlation values. For fluency assessment, deep learning models like wav2vec were used to extract audio features, and an SVM classifier trained on a fluency-focused corpus categorized the samples. The outcomes highlight the complexities of evaluating such phenomena, with variability depending on the specific type of disfluency detected.

Index Terms: speech disorders, pronunciation assessment, disfluency detection, Down syndrome

1. Introduction

Identifying elements of pathological speech is crucial for diagnosing and treating various speech disorders. Accurate identification of the specific problem that the speaker has (from a long list of potential issues, such as dysarthria, stuttering, cluttering ...) and the symptoms that indicates a possible pathology (e.g. speech blocks, changes in pronunciation, repetitions ...) helps develop targeted intervention strategies that improve the quality of life for people with language disorders. However, annotating audio to identify these problems is not a common practice in speech therapy. This has led to a shortage of linguistic resources needed to train automatic evaluation systems or automatic identification of language pathology.

The are some studies that annotated corpora of pathological speech. The authors in [1] use the GRBAS scale to annotate a speech corpora and correlate the scores with acoustic parameters – global Grade of dysphonia (G), Roughness (R), Breathiness (B), Asthenicity (A), and Strain (S) and the study described in [2] uses this scale for training automatic evaluation systems. There are some corpora focused on fluency disorders [3, 4, 5]. In [6] stuttering is cross-analyzed with the perception of good communication skills. In [7] the quality of oral productions of individuals with Down syndrome (DS) was annotated, and a comprehensive review of the currently available corpora was presented.

Although the Goodness of Pronunciation (GoP) method is commonly used to assess the pronunciation of non-native (L2) speech [8], several studies have also demonstrated its effectiveness in evaluating speech disorders. For example, the GoP measure has been applied to disordered speech from speakers with unilateral facial palsy [9], children with apraxia of speech [10, 11], and children with cleft lip/palate [12]. Additionally, GoP scores have been utilized to predict comprehensibility ratings in patients with neurological and anatomic speech disorders [13], as well as for the automatic speech intelligibility assessment of dysarthric speech [14].

Related with automatic stuttering detection, there are some reviews focused on applying machine learning approaches to build automatic stuttering identification systems (ASIS) [15, 16]. Statistical methods (HMM, SVM, KNN) as well as neural networks (ANNs, RNNs, LSTMs, CNNs) have been used, obtaining different results depending on the corpus and the methodology applied in each study. However, the study of stuttering in people with Down syndrome using these methodologies is limited, taking into account that a high rate of stuttering occurrence in individuals with Down syndrome, independent of assessors, has been detected compared to typically developing individuals [17].

Since people with DS have unique characteristics that make training automatic systems challenging, an annotated corpus with enriched information is necessary. We present the annotation rubric and the initial results of this effort on a speech corpus called PRAUTOCAL [7], which consists of a large number of utterances from Spanish speakers with Down syndrome. The corpus was recorded using an educational game.

The structure of the paper is as follows. We begin by describing the annotation of the corpus, which encompasses phonetic and fluency dimensions. Following this, we present the phonetic experiments using the GoP metric. Next, we detail the stuttering experiments aimed at identifying five types of disfluencies. Finally, we conclude with a discussion, conclusions, and suggestions for future work.

2. Corpus annotation

The PRAUTOCAL corpus is a corpus of Spanish speakers with Down syndrome from the northern/central Iberian Peninsula, which allows the analysis of specific aspects of the speech of individuals with Down syndrome. It also includes comparable recordings of typically developing (TD) users for reference. So far, the corpus has been used for prosodic studies. In the work described in this paper we aim to also use the corpus to evaluate phonetic pronunciation and fluency. The corpus was collected in six recording campaigns and contains 90 speakers, with 4,175 audio files and a total audio duration of approximately 3 hours and 47 minutes. The corpus is balanced in terms of gender (49 men and 41 women) and speaker type (50 individuals with intellectual disabilities and 40 typically developing speakers). The age range for both speaker types is also similar, between 13 and 42 years for speakers with DS and between 6 and 68 years for TD speakers. The corpus also includes the tran-

10.21437/IberSPEECH.2024-6

scription of the utterances, which has been used for automatic phonetic segmentation of the recordings. A detailed description of the PRAUTOCAL corpus can be found in [7].

2.1. Annotation process

The annotation process was made using a rubric, that was elaborated in a three steps designing approach. First, a draft was elaborated by the authors and validated by expert linguists and therapists. Then, it was used to conduct a controlled evaluation campaign in which a reduced number of samples and experts participated. Finally, from the conclusions in this step and after analyzing the main sources for inter-rater disagreement, a final version was delivered and used by an evaluator which had not participated in previous steps. We have formulated the evaluation rubric that encompasses three dimensions: phonetics, fluency, and prosody. In the experiments described in this paper the prosody dimension is not used.

A speech therapist with several years of experience has annotated the corpus. A total of 2,084 utterances (12,303 words in the phonetic part) have been annotated, all of them produced by people with Down syndrome. The annotation was made using a web page in which the speech therapist could listen to the phrase to be tagged as many times as she wanted and in which the errors and the assessment had to be indicated.

2.2. Phonetics

Regarding phonetics, the evaluator marked articulation errors at the segmental level in words, including substitution, omission, distortion, and addition (SODA) [18, 19, 20]:

- Substitution: one phoneme is substituted by another.
- Omission: a phoneme is deleted.
- Distortion: a phoneme is not replaced by another, but is not articulated according to what is expected.
- Addition: a phoneme is added to the word.

In addition, the evaluator must assess the overall phonetics quality of the sentence using the following three levels:

1. Speech with frequent errors or distortions of sounds.
2. Errors are sporadic, appearing in some situations but not in others.
3. Pronunciation is correct without obvious pronunciation errors.

"Sporadic" errors occur in 25% or less of the words within the utterance, while "frequent" errors appear in more than 25% of the words.

Figure 1 shows the distribution of the phonetic and fluency assessment. The distribution of the phonetic scores is quite balanced. The number of words with errors and the percentages out of the total number of words are shown in Table 1.

2.3. Fluency

Concerning fluency, for each category of fluency errors, the evaluator must indicate whether each deviation occurs once, more than once, or not at all:

- Block: an involuntary pause before a word or within a word, due to physiological reasons. Sometimes, a breath can be perceived, but in most cases, there is no breath or it is inaudible. These pauses are not linguistically motivated.
- Prolongation: an involuntary lengthening or prolongation of a syllable or sound.
- Repetition of sounds/syllables.

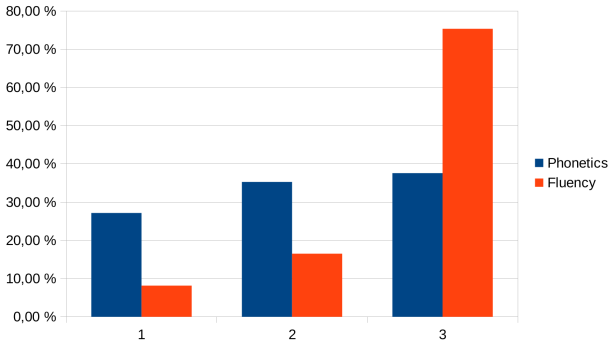

Figure 1: *Distribution of scores for phonetic and fluency quality (1: frequent errors; 2: sporadic errors; 3: without errors).*

- Repetition of words/phrases.
- Interjection: filler words, which are used to buy time to find the right words to continue speaking. They are usually employed when the speaker has difficulties to pronounce a specific word, and fillers provide time to think of an alternative word that is easier to pronounce.

Moreover, the evaluator should determine the overall quality of the fluency of the sentence using the following three levels:

1. Speech with frequent fluency errors.
2. Speech with sporadic fluency errors.
3. Speech without fluency errors.

"Sporadic" means that the number of errors is 25% or less of the number of words in the utterance, while "frequent" means that the number of errors is more than 25% of the number of words in the utterance.

As shown in Figure 1, the distribution of fluency scores is clearly unbalanced. Table 2 shows the number of utterances without errors, one error, or more than one error for each type of fluency error.

Table 1: *Errors found in the phonetic part of the analysis. The errors are marked at word level.*

Substitution	526 (4.3%)
Omission	1,108 (9.0%)
Distortion	3,056 (24.8%)
Addition	318 (2.6%)

Table 2: *Errors in the fluency part.*

	None	**One**	**Multiple**
Blocks	1,599	356	129
Prolongations	1,922	106	56
Sound repetitions	1,797	225	62
Word repetitions	1,843	162	79
Interjections	2,003	69	12

3. Experiments

In this section, we describe the preliminary experiments conducted to compute the baseline classification results. These ex-

periments use common techniques for phonetic and fluency assessment.

3.1. Phonetic experiments

In this experiment we have used Goodness of Pronunciation (GoP) [8] for automatic assessment of phonetic quality. GoP is a measure of the degree of similarity between produced and canonical pronunciation of phonemes.

There are several GoP methods described in the bibliography. The first method described was GMM-GoP [8], which, for a phone p is defined as an averaged log probability across the phone duration:

$$\text{GMM-GoP}(p) = \frac{1}{|F|} \sum_{f \in F} log \frac{e^{L^f(p|f)}}{\sum_{q \in Q} e^{L^f(q|f)}} \quad (1)$$

where the duration of phone p in frames is F; $P^f(p|f)$ is the frame-wise phone probability and $L^f(p|f)$ its logits; Q is the total phone set.

With the development of neural networks (NNs), variants of GoP which employ probabilities from the state-of-the-art neural networks have been suggested [21]:

$$\text{NN-GoP}(p) = \log \bar{P}(p|F) - \max_{q \in Q} \log \bar{P}(q|F) \quad (2)$$

$$\bar{P}(p|F) = \frac{1}{|F|} \sum_{f \in F} P^f(p|f) \quad (3)$$

Finally, DNN-GoP [21] normalizes the phone probability with the phone prior:

$$\text{DNN-GoP}(p) = \frac{\bar{P}(p|F)}{P(p)} \quad (4)$$

In order to calculate the GoP measures, we employ a self-supervised learning approach to extract posterior probabilities from the widely adopted cross-lingual wav2vec 2.0 XLS-R model [22], in line with recent research [23, 14]. The fine-tuning process uses the Common Phone dataset [24] to train a linear phone prediction head on top of the wav2vec model. Notably, this linear phone prediction head is incorporated above the convolutional layer rather than the transformer layer. This design choice serves to reduce the computational complexity of the model while preserving important phonetic characteristics in the convolutional features. The optimization is performed using the AdamW optimizer [25], employing a default learning rate of 0.001, and this process is repeated for four epochs.

The acoustic model has been trained on a collection of speech samples of typical healthy speakers drawn from the Common Phone dataset. This dataset was specifically chosen for its comprehensive coverage of phonemes and detailed phonetic annotations, making it a prime choice for encompassing a broad spectrum of Spanish phonemes. The Common Phone dataset is noteworthy for its gender balance and multilingual content, spanning six different languages. With over 11,000 speakers contributing to it, the dataset boasts approximately 116 hours of recorded speech.

We evaluated the efficacy of the model using the PRAUTOCAL corpus. First, in order to obtain the segmentation, Montreal Forced Aligner (MFA) [26] is employed to extract phoneme-level alignments in the PRAUTOCAL corpus. Then, we calculated the average GoP score for each utterance and determined its correlation with the intelligibility score. We used the Kendall Rank Coefficient τ to measure the correlation between the average GoP scores and the phonetic quality assessments. Two different configurations were used:

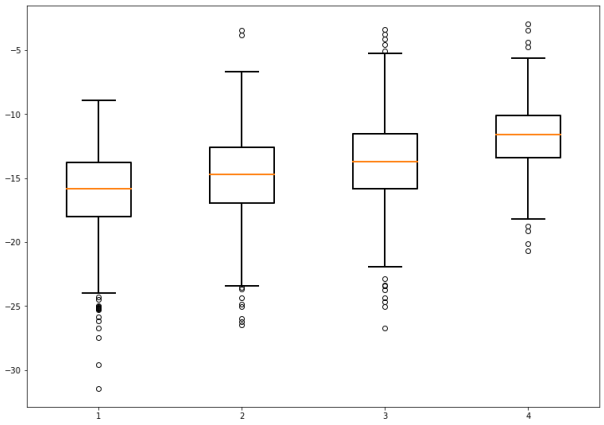

Figure 2: *Distribution of the GMM-GoP for the four different score levels.*

- 3-level: we used 2,084 utterances from people with DS, evaluated with the 3-level assessment (1, 2 or 3) as described in section 2.
- 4-level: we added 700 randomly chosen utterances of typically developing speakers from the PRAUTOCAL corpus. As they are supposed to be correctly pronounced utterances, the assessment was set to 4.

Results are shown in Table 3, highlighting the performance of different configurations. In the 3-level setup, the best outcome is achieved by GMM-GoP, with a score of 0.1983. However, when employing the 4-level configuration, the best result is obtained by DNN-GoP, attaining a score of 0.3667. Notably, the 4-level configuration consistently outperforms the 3-level counterpart in all scenarios. Figure 2 offers an insight into the distribution of GMM-GoP measurements for scores 4, 3, 2 and 1 (similar graphics are observed for NN-GoP and for DNN-GoP). The highest GoP values are obtained for score 4 (TD speakers), as their speech is expected to be very similar to the canonical pronunciation of phonemes, while the lowest values are associated with score 1. Scores 2 and 3 exhibit intermediate values between score 1 and score 4. In summary, the results align with anticipated trends, despite the somewhat modest correlation values.

Table 3: *Kendall's rank coefficient between GoP measures and phonetic evaluation scores of the rubric.*

	3-level	**4-level**
GMM-GoP	0.1983	0.3345
NN-GoP	0.1635	0.3176
DNN-GoP	0.1410	0.3667

3.2. Stuttering experiments

In the stuttering experiments, we employed an approach similar to the one used in [27]. As a reference corpus, we utilized the KSoF corpus [5], which is a corpus containing 5,597

3-second recordings obtained from stuttering therapy sessions. These recordings were labeled by three annotators for five types of disfluencies: blocks, prolongations, sound repetitions, word repetitions, and interjections, in addition to other additional labels that we have not considered in this study. The distribution of labels in this corpus is highly unbalanced.

As feature extractor, we used the base model of wav2vec [28], which is a model trained in an unsupervised manner with 960 hours of unlabeled speech data from the LibriSpeech corpus [29]. This model was adapted for automatic speech recognition using transcriptions of the same audio. For each labeled audio, we extracted 768-dimensional speech representations every 20ms, and the mean was calculated to obtain representations for the entire audio. We used SVMs as classifiers because they have demonstrated effectiveness in delivering robust results even when working with a limited set of samples. The classification is a binary tasks of one specific disfluency against all other samples.

To obtain the optimal hyper-parameters of the classifier for each type of disfluency, we conducted classification using the embedding extracted from each of the layers provided by the model, employing principal component analysis (PCA) to obtain features that explain a minimum of 0.9 variance. The kernel parameters were selected from {0.1, 0.01, 0.001, 0.0001, 0.00001}, and the penalty parameter of error C was chosen from {1, 10, 100, 1000}. From the results obtained, we selected the parameters and layer that achieved the best results (F1 score) for each type of disfluency. Finally, an SVM classifier was trained using all the samples with the parameters obtained in the previous phase.

The classifier trained on the KSoF corpus was used to classify samples from our corpus and the results are shown in Table 4. As can be seen in Table 2, the labels used for the fluency evaluation are similar to those used in the KSoF corpus for the purpose of comparison. However, it is important to note that both the evaluators and their interpretation of the evaluation criteria may differ. As a result, we cannot establish an exact correspondence between the labels in both corpora, even if the disfluency to be detected is the same. This variation can impact on the classification results. Additionally, just like in the KSoF corpus, the labels for different disfluencies are highly unbalanced. In the PRAUTOCAL corpus, the identification of disfluencies is not binary, as the evaluator has three levels of assessment for each type of disfluency: no disfluency, one disfluency, or two or more disfluencies. To convert this assessment into a binary format, it is considered that there is disfluency if the label indicates one or more disfluencies.

The results in Table 4 are obtained by using the features of each layer of wav2vec and selecting the maximum F1 score for each label in disfluency detection/no disfluency detection. For blocks, an F1 score of 0.42/0.46 is obtained; for prolongations, an F1 score of 0.16/0.48; for sound repetitions, an F1 score of 0.28/0.66; for word repetitions, an F1 score of 0.13/0.74; and for interjections, an F1 score of 0.07/0.52 is obtained.

Table 4: *F1 score of each disfluency label. KSoF column shows the best result in the optimization process (disfluency detection).* PRAUTOCAL *column shows the results obtained with the classifier trained with the KSoF samples and tested with the* PRAUTOCAL *corpus. Id means disfluency detection and nId means no disfluency detection.*

	KSoF	**PRAUTOCAL (id/nid)**
Blocks	0.56	0.42 / 0.46
Prolongations	0.58	0.16 / 0.48
Sound repetitions	0.36	0.28 / 0.66
Word repetitions	0.44	0.13 / 0.74
Interjections	0.16	0.07 / 0.52

4. Discussion

We performed a series of experiments using the annotations obtained from the rubric. Regarding phonetic experimentation, we conducted baseline machine learning trials, which yielded results indicating a moderate correlation between the GoP values and assessment scores. This outcome was achieved through a straightforward approach of averaging GoP across all phonemes within an utterance. While this serves as an initial step for future research, there is potential to enhance results through the implementation of novel classification techniques. Moreover, it is imperative to conduct an in-depth examination of GoP values for each individual phoneme. Individuals with Down syndrome often face greater challenges in articulating specific phonemes while displaying relative proficiency in others. For instance, in the context of the Spanish language, individuals with DS may find it especially challenging to master the /rr/, /r/, and /z/ phonemes [30].

The results in Table 4 highlight the challenge of evaluating different disfluencies using automatic classifiers. The best results are obtained for blocks and prolongations, although they are still far from the results achieved in the KSoF corpus. Nonetheless, F1 values for the absence of disfluency are higher for all types of disfluencies. This is particularly important during therapy sessions, as an inaccurate classification of disfluency can lead to patient frustration. Labeling the corpus remains a complex task, even when the same disfluencies are included, because different evaluators may interpret disfluencies differently. In the KSoF corpus used to train the classifier, there is not a high level of agreement among evaluators [5], which demonstrates the challenge of applying consistent criteria when evaluating speech disfluencies.

5. Conclusions

This paper outlines the phonetic and fluency annotation of the PRAUTOCAL corpus, which includes speech data from Spanish speakers with Down syndrome. To evaluate the effectiveness of these annotations in automatic classification systems, two exploratory experiments are presented. Although the initial results can be improved upon, having a corpus annotated for individuals with Down syndrome in various aspects of speech holds the potential to enhance these outcomes. This paves the way for developing automatic systems that could benefit speech therapy for individuals with Down syndrome.

As future work, it is crucial to incorporate new evaluators to improve the labeling process. Additionally, fine-tuning the models used and implementing other deep learning techniques could further enhance the results.

6. Acknowledgements

This work was carried out in the Project PID2021-126315OB-I00 that was supported by MCIN / AEI / 10.13039/501100011033 / FEDER, EU.

7. References

[1] G. Demenko, A. Obrebowski, A. Pruszewicz, B. Wiskirska-Woznica, P. Swidzinski, and W. Wojnowski, "Suprasegmental analysis for complex quality assessment in pathological voices," in *Speech Prosody 2004, International Conference*, 2004.

[2] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the GRBAS scale," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 2478–2481.

[3] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6798–6802.

[4] N. B. Ratner and B. MacWhinney, "Fluency bank: A new resource for fluency research and practice," *Journal of fluency disorders*, vol. 56, pp. 69–80, 2018.

[5] S. Bayerl, A. Wolff von Gudenberg, F. Hönig, E. Noeth, and K. Riedhammer, "KSoF: The kassel state of fluency dataset – a therapy centered dataset of stuttering," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, 2022, pp. 1780–1787.

[6] D. Werle and C. T. Byrd, "Professors' perceptions and evaluations of students who do and do not stutter following oral presentations," *Language, Speech, and Hearing Services in Schools*, vol. 53, no. 1, pp. 133–149, 2022.

[7] D. Escudero-Mancebo, M. Corrales-Astorgano, V. Cardeñoso-Payo, L. Aguilar, C. González-Ferreras, P. Martínez-Castilla, and V. Flores-Lucas, "Prautocal corpus: a corpus for the study of Down syndrome prosodic aspects," *Language Resources and Evaluation*, vol. 56, no. 1, pp. 191–224, 2022.

[8] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[9] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, "The goodness of pronunciation algorithm applied to disordered speech," in *Proc. Interspeech 2014*, 2014, pp. 1463–1467.

[10] M. A. Shahin, B. Ahmed, J. X. Ji, and K. J. Ballard, "Anomaly detection approach for pronunciation verification of disordered speech using speech attribute features." in *INTERSPEECH*, 2018, pp. 1671–1675.

[11] M. Shahin and B. Ahmed, "Anomaly detection based pronunciation verification approach using speech attribute features," *Speech Communication*, vol. 111, pp. 29–43, 2019.

[12] V. C. Mathad, T. J. Mahr, N. Scherer, K. Chapman, K. C. Hustad, J. Liss, and V. Berisha, "The impact of forced-alignment errors on automatic pronunciation evaluation." in *Interspeech*, 2021, pp. 1922–1926.

[13] L. Fontan, T. Pellegrini, J. Olcoz, and A. Abad, "Predicting disordered speech comprehensibility from goodness of pronunciation scores," in *Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2015) satellite workshop of Interspeech 2015*, 2015.

[14] E. J. Yeo, K. Choi, S. Kim, and M. Chung, "Speech Intelligibility Assessment of Dysarthric Speech by using Goodness of Pronunciation with Uncertainty Quantification," in *Proc. Interspeech 2023*, 2023, pp. 166–170.

[15] L. Barrett, J. Hu, and P. Howell, "Systematic review of machine learning approaches for detecting developmental stuttering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1160–1172, 2022.

[16] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "Machine learning for stuttering identification: Review, challenges and future directions," *Neurocomputing*, vol. 514, pp. 385–402, 2022.

[17] S. Hokstad and K.-A. B. Næss, "Stuttering in individuals with Down syndrome: a systematic review of earlier research," *Frontiers in Psychology*, vol. 14, p. 1176743, 2023.

[18] N. Sreedevi and A. Mathew, "Articulation errors in Malayalam speaking children with hearing impairment who use digital hearing aids: an exploratory study," *Journal of Hearing Science*, vol. 12, no. 2, pp. 49–59, 2022.

[19] J. Gallardo and J. Gallego, "Alteraciones de la articulación: dislalias," *JR Gallardo y JL Gallego, Manual de Logopedia Escolar. Un enfoque práctico*, pp. 171–220, 1995.

[20] F. Susanibar, A. Dioses, and J. Tordera, "Principios para la evaluación e intervención de los trastornos de los sonidos del habla–TSH," *Susanibar F, Dioses A, Marchesan I, Guzmán M, Leal G, Guitar B, Junqueira Bohnen. Trastornos del Habla. De los fundamentos a la evaluación. Madrid. EOS*, 2016.

[21] 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, 2015.

[22] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[23] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection." in *Proc. Interspeech 2021*, 2021, pp. 4428–4432.

[24] P. Klumpp, T. Arias, P. A. Pérez-Toro, E. Noeth, and J. Orozco-Arroyave, "Common phone: A multilingual dataset for robust acoustic modelling," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, 2022, pp. 763–768.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[26] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[27] S. P. Bayerl, D. Wagner, E. Noeth, and K. Riedhammer, "Detecting Dysfluencies in Stuttering Therapy Using wav2vec 2.0," in *Proc. Interspeech 2022*, 2022, pp. 2868–2872.

[28] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[30] J. Sánchez Marin, "Comparación del desarrollo fonético-fonológico de niños con síndrome de Down y desarrollo típico: influencia de los aspectos madurativos y cognitivos," Ph.D. dissertation, Universidad de Murcia, 2014.