WawaSimi: Classification Techniques for Phonological Processes Identification in Children from 3 to 5 Years Old

Braulio Baldeon, Renzo Ravelli and Willy Ugarte[a]
Universidad Peruana de Ciencias Aplicadas (UPC), Lima, Peru

Keywords: Phonological Processes, Language Development Level, Language Acquisition, Machine Learning.

Abstract: In the context of the pandemic we are living in, most of the interactions between kindergarten-aged children has decreased, meaning that their language development might be slowed down. Our work presents a machine learning-based method for the classification of phonological processes and a corpus with a total of 3,324 audios, being 40% of them audios with a correct pronunciation, and the remaining 60% wrong. One of the main problems encountered when trying to perform children speech recognition in Spanish, is, to the best of our knowledge, the lack of a corpus. 329 audios were collected from 20-30 years old adults and a voice conversion technique was applied in order to generate the required audios for the corpus construction. A modified AlexNet was trained to classify if a word was correctly pronounced, if the audio was classified as mispronounced, it goes through a second AlexNet to classify which kind of Phonological Process is found in the word.

1 INTRODUCTION

The COVID-19 pandemic has brought many unintended consequences for society, such as a negative effect on communication, especially in the pediatric population, as preventative practices like wearing masks, social distancing, and virtual classrooms have a direct impact in the early-ages language development since social interaction is essential for this process.

Social Distancing and Virtual Classrooms have affected children from having meaningful in-person interactions with peers; this is a crucial component of pragmatic development, which include conversational skills as turn talking and understanding the meaning behind the other speaker's words. Masks also obscure social cues provided by facial expressions (Charney et al., 2021).

In 2021's first quarter a survey was conducted by the Education Endowment Foundation, which consisted on collecting data from 58 primary schools across England, showed that 76% of the pupils starting school in September 2020 needed more communication support than in previous years; 96% of schools claimed to be concerned about their pupils' speech-and-language development and 56% of parents were concerned about their child starting school following the lockdown in the spring and summer[1].

Through the pandemic, parents have done an amazing job to keep their children safe and healthy. But this has reduced children's exposure to new vocabulary, as they are not able to hear and learn words that family might use when the visit the extended family. Mask wearing has also had an impact in language development, as in school and pre-school, children may struggle to differentiate between similar sounds such as "p" and "t", when their teacher is wearing a mask. This can impact on a child's speech development or their phonological awareness; masks also obscure facial expression, which contributes to how we understand the meaning behind words[2].

At the beginning of August 2020, the Peruvian Ministry of Health presented a statement regarding language and speech disorders during this pandemic. In this statement, it was noted that children are one of the most vulnerable population in the context we live in, and there are reported cases of minors who develop difficulty in articulating sounds, suffer alterations in the fluency of speech and have problems in the acquisition of expressive and / or comprehensive language.

[a] https://orcid.org/0000-0002-7510-618X

[1] "Lockdowns hurt child speech and language skills - report" - BBC

[2] "How lockdown has affected children's speech – and what parents can do to help" - The Conversation

Baldeon, B., Ravelli, R. and Ugarte, W.
WawaSimi: Classification Techniques for Phonological Processes Identification in Children from 3 to 5 Years Old.
DOI: 10.5220/0010991300003182
In *Proceedings of the 14th International Conference on Computer Supported Education (CSEDU 2022) - Volume 2*, pages 147-154
ISBN: 978-989-758-562-3; ISSN: 2184-5026

The Ministry's representative noted that confinement causes children to have less social interaction, which has a negative impact on increasing their vocabulary and communication skills. She also noted that by the end of 2020's second quarter, Community Mental Health Centers treated 6,846 cases of language problems across the country[3].

In order to address this problem, some researches has been done, some of them on the classification of mispronunciation of words, using different machine-and-deep learning methods in order to do so. As an example, for the Arabic alphabet, a comparison between classifiers was conducted; the results showed that using a transfer-learning approach, taking advantage of an AlexNet outperformed other methods, leading to a 99%+ accuracy on the task (Terbeh and Zrigui, 2017). Due to this results, the transfer-learning approach was selected in order to perform the mispronunciation classification.

On the other hand, in order to conduct the identification of phonological processes in Brazil, considering different types and not only one, additional data was provided by speech therapists (Franciscatto et al., 2021). As we don't have access to this data for Spanish, we decided to take the same approach as for the mispronunciation classification.

Our contributions are as follows:

- We have built a synthetic dataset that mimics both the pitch and speaking rate of 3-5 years old Spanish-speaking children.
- We propose a neural network architecture to detect mispronounced words based on their Mel-spectrogram.
- We propose a neural network architecture to identify phonological processes (Structure of the syllable and the word, Assimilation or Substitution) in mispronounced words.

This paper is organized as follows. Section 2 discusses related work. Section 3 explains the context of the problem and our main contributions in greater detail. In section 4 we present the results obtained from the experimentation and we conclude in section 5.

2 RELATED WORKS

In (Terbeh and Zrigui, 2017), the authors propose the usage of 3 different neural networks (CNN, pre-trained AlexNet and BLSTM) to classify which letter of the Arabic alphabet has been pronounced and its pronunciation quality. Additionally, they apply data augmentation through pitch modification. After removing noise and silences from each audio file, they used Mel spectrograms, with three color channels, as input data for the first two networks; and spectral characteristics manually extracted for the last network. Our proposal uses a pre-trained AlexNet to detect bad pronunciation and PP's; using MFCC heat maps with a single color channel. On the other hand, being in a scenario with no data available, we use the modification of pitch and speaking rate to generate synthetic data instead of just increasing it.

In (Nazir et al., 2019), The authors propose a method based on characteristics extracted by a pre-trained AlexNet and a method based on transfer learning to detect mispronunciation in people who learn Arabic as a second language. In the first method, they used the characteristics extracted by AlexNet from Mel spectograms as input data from three different models (KNN, SVM and a neural network). While in the second method, they used the Mel spectrograms, with three color channels, as input data for a pre-trained AlexNet. Unlike them, we propose the use of MFCC heat maps with a single color channel as input data for a pre-trained AlexNet to detect mispronunciation in children whose native language is Spanish.

In (Franciscatto et al., 2021), the authors presents a method that consists on a Pronunciation Recognition module, using a Decision Tree model, and a Phonological Process Recognition module, that uses a Random Forest model using a matrix given by an expert with the probabilities of different Phonological Processes in each phoneme of a set of words. Inspired by this structure, our proposal uses a transfer-learnt AlexNet to recognize mispronunciation and another AlexNet for the identification of Phonological Processes.

In (Shahnawazuddin et al., 2020), the authors present a prosody-modification-based data augmentation method for automatic speech recognition. This is performed locating the Glottal Closure Instants (GCIs) to use them as anchors to modify the Speaking-Rate and Pitch from the audios; after this process, the audio is constructed again with the new pitch and speaking-rate. While our method is very similar, as we perform the modification of Speaking-Rate and Pitch, we don't calculate the GCIs, and modify the Speaking-Rate and Pitch directly.

[3]Ministry of Health warns of an increase in language disorders in children due to the emergency (in spanish)

Table 1: Examples of errors (Pavez and Coloma, 2017).

(a) Syllable structure errors.

Structure of the syllable	Error in spanish			(Word in english)
Consonant-group reduction	/p_áto/	for	/pláto/	(plate)
Diphthong reduction	/á_to/	for	/áuto/	(car)
Coda suppression	/pa_talón/	for	/pantalón/	(pants)
Coalescence	/kén/	for	/tren/	(train)
Omission of unstressed elements	/pósa/	for	/mariposa/	(butterfly)
Addition of phonemes or syllables	/níndio /	for	/indio/	(Indian)
Inversion of phonemes or syllables	/uáto/	for	/áuto/	(car)

(b) Assimilation errors.

Assimilation	Error in spanish			(Word in english)
Identical	bubánda/	for	/bufánda/	(scarf)
Labial	/plátamo/	for	/plátano/	(banana)
Dental	/madípósa/	for	/maripósa/	(butterfly)
Velar	/gufánda/	for	/bufánda/	(scarf)
Nasal	/anfómbra/	for	/a/fómbra/	(carpet)
Syllabic	/lilikóptero/	for	/elikóptero/	(helicopter)

(c) Substitution errors.

Substitution	Error in spanish			(Word in english)
Posteriorization	/ekifísio/	for	/edifício/	(building)
Frontalization	/buánte/	for	/guánte/	(glove)
Stopping	/póka/	for	/fóka/	(seal)
Fricativization	/marifósa/	for	/maripósa/	(butterfly)
Semiconsonantization of liquid phonemes	/tjen/	for	/tren/	(train)
Within-category liquid substitution	/kape/usita/	for	/kaperusita	(Little Red Riding Hood)

3 IDENTIFYING AND CLASSIFYING PHONOLOGICAL PROCESSES

This section presents our method by introducing core notions and developing them into a model proposal.

3.1 Preliminary Concepts

Now, we describe the main concepts for our proposal.

3.1.1 Phonological Processes

Also known as speech pattern errors, phonological processes (PP's) are mental operations that apply on speech to substitute a class of sound or sound sequences that are difficult to the speech capacity of the individual, for an alternative class identical but lacking the difficulty property (Franciscatto et al., 2021).

Structure of the Syllable and Word. In the processes related to the structure of the syllable and word,

the emissions are simplified by bringing them closer to the basic syllabic structure (consonant + vowel). Some structure of the syllable and word phonological processes are presented in Table 1 along with their corresponding sub-class (Coloma et al., 2010).

Assimilation. Assimilation processes are strategies by which a phoneme is replaced to make it the same or similar to another present in the word (Coloma et al., 2010). Some examples of this kind of phonological processes are shown in Table 1b.

Substitution. In substitution processes, phonemes belonging to one class are exchanged for members of another phonemic class (Coloma et al., 2010). Some examples of this kind of phonological processes are shown in Table 1c.

3.1.2 Deep Learning

Deep learning uses a multi-layered cascade of non-linear processing units for feature extraction and transformation. The lower layers near the data entry learn simple features; in contrast, the upper layers learn more complex characteristics derived from the characteristics learned by the lower layer as shown in Fig. 1. In recent years, deep learning has produced cutting-edge results in many domains such as computer vision, speech recognition, and NLP (Zhang et al., 2018).

Figure 1: Characteristics learned by lower and upper layers (Zhang et al., 2018).

Convolutional Neural Network. Is a deep learning model used to analyze and learn from data or information that has a mesh or grid pattern. These models generally have three different types of layers: convolutional, grouping, and fully connected. The first two perform the feature extraction process autonomously through convolutions and the third locates these features in the final output. A convolution is a type of linear operation used to extract features; basically, a small matrix of numbers, called a kernel, is applied to the input data, tensor (a matrix of numbers) as shown in Fig 2. The tensor and the kernel are used to calculate a scalar product for each element of the tensor; the result is used as the output value in the corresponding position of the characteristic map (Khan et al., 2020).

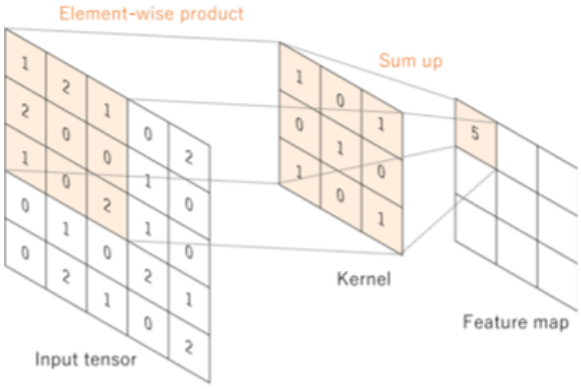


Figure 2: Convolution example (Khan et al., 2020).

Transfer Learning: This method aims to provide a framework for using previously acquired knowledge to solve new but similar problems much more quickly and effectively. This method has been inspired by the fact that human beings can use previously acquired knowledge to solve new but similar problems; for example, learning to recognize apples can help us to recognize pears, or learning to play the electronic organ

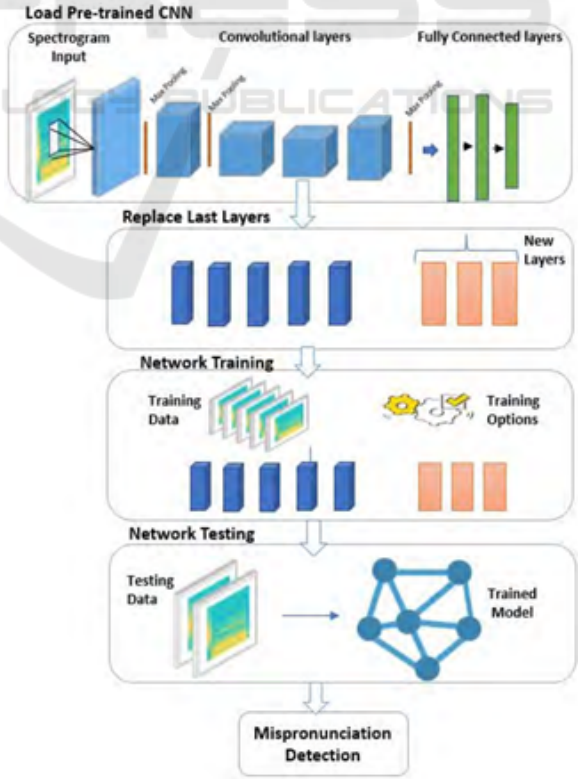


Figure 3: Workflow for applying transfer learning with CNN (Lu et al., 2015).

Table 2: Target words to identify phonological processes.

Correctly pronounced		Incorrectly pronounced
/alfómbra/	(carpet)	/a_fómbra/
/áuto/	(car)	/á_to/, /uáto/
/bufánda/	(scarf)	/bubánda/, /gufánda/
/kaperusita/	(Little Red Riding Hood)	/kaperutusita/, /kape_usita/
/edifício/	(building)	/ekifísio/
/fóka/	(seal)	/póka/
/guánte/	(glove)	/buánte/
/elikóptero/	(helicopter)	/lilikóptero/
/indio/	(Indian)	/níndio/
/mariposa/	(butterfly)	/pósa/, /madípósa/, /marifósa/
/pantalón/	(pant)	/pa_talón/
/plátano/	(banana)	/plátamo/
/pláto/	(plate)	/p_áto/
/teléfono/	(phone)	/tenéfolo/
/tren/	(train)	/t_en/, /kén/, /tjen/

Table 3: Number of generated audio files.

	Correctly pronounced	Incorrectly pronounced	Total
Male (10)	$10 \times 15 \times 6 = 900$	$(10 \times 22 - 1) \times 6 = 1,314$	2,214
Female (5)	$5 \times 15 \times 6 = 450$	$5 \times 22 \times 6 = 660$	1,110
Total (15)	1,350	1,974	3,324

can make learning the piano easier (Lu et al., 2015). The way we apply this method in a convolutional network context is shown in Fig 3.

3.1.3 Voice Conversion

Voice Conversion (VC) is an area of speech processing that deals with the conversion of the speaker's perceived identity. VC aims to modify the non-linguistic or paralinguistic information of speech while preserving the linguistic information; that is, the voice signal emitted by a first speaker, the source speaker, is modified so that it sounds as if it were spoken by a second speaker, the target speaker (Popa et al., 2012) as shown in Fig. 4.

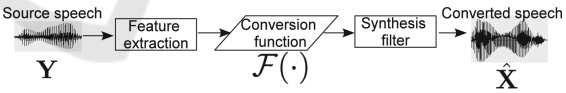


Figure 4: Voice conversion example (Wu and Li, 2014).

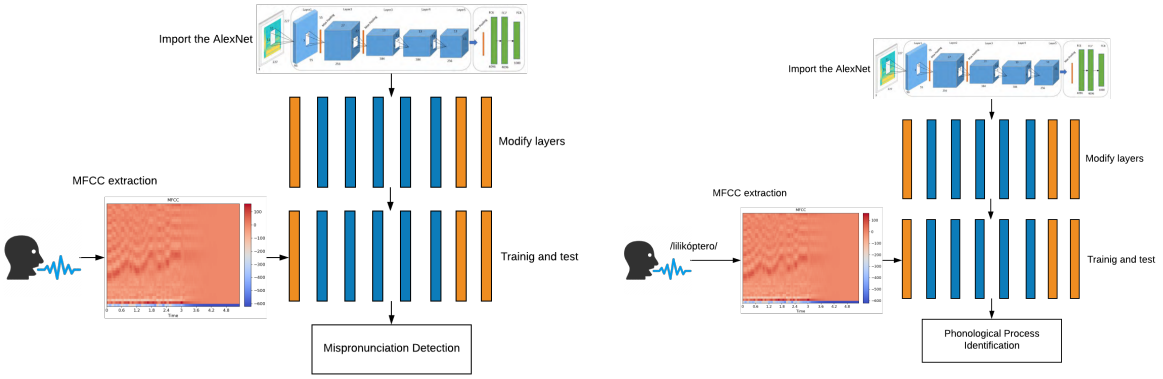
3.2 Method

In this section we will explain our main contributions: creation of a synthetic data using voice conversion, applying transfer learning in a CNN to detect mispronounced words and the same process to identify PP's.

3.2.1 Synthetic Dataset

Due to the fact that, to the best of our knowledge, there was not a corpus of Spanish children's speech available so far, we had to use Voice Conversion to generate synthetic audios that imitate the voice of 3-to-5 years old children.

This process consists of modifying the pitch and speaking rate of audios collected from people between 20 and 25 years old. In order to perform this modification, we considered an average pitch of 309 Hz for 3-years-old children (Schuckman, 2008),



(a) Mispronunciation detection process workflow.

(b) Phonological process identification workflow.

Figure 5: Processes Workflows.

255Hz for 4 years and 253 Hz for 5 years (Cappellari and Cielo, 2008). Likewise, we use a speaking rate of 5 or 7 phonemes per second for all three ages (Shahnawazuddin et al., 2020), resulting in 6 combinations.

A group of 15 persons were asked to record their voices while pronouncing a set of words commonly used to identify Phonological Processes (Popa et al., 2012). This set of words consist on 15 words, and 22 mispronunciation of this 15 words as seen in Table 2; giving a total of 37 audios.

After collecting the audios, they were processed detecting and removing silences from the beginning and end of them in order to get the audios with only the words; these processed files were modified (pitch and speaking rate) using the Python librosa library. From this process we obtained a synthetic corpus of 3324 audio files, because one person only recorded 36 words, as shown in the Table 3.

3.2.2 Mispronunciation Detection

To detect whether a word was pronounced correctly or incorrectly, we used the Mel frequency cepstral coefficients (MFCC) of the generated audio files and a previously trained convolutional neural network (AlexNet). We use AlexNet available in the PyTorch library, which was trained on the ImageNet dataset and can classify between 1000 different classes.

The detail of our sequence of steps is shown in Fig. 5. Firstly, we extract the MFCC using the librosa library and resize its heat map to $227 \times 227 \times 1$. Secondly, we import the Alexnet and modify the first convolutional layer, such that it has the same input size as the heat maps, and the last two layers completely connected, so that it has an output according to our number of classes (correctly and incorrectly pronounced). Thirdly, we divided the corpus into training (70%), test (30%) and validation (222 audios); for which we employ two approaches "Known Speaker" and "Unknown Speaker".

- **Known Speaker (KS):** the partition is done without taking into account the source of the audio. In such a way that the audios emitted by the same person can be in the training and test partitions.
- **Unknown Speaker (US):** the partition is carried out taking into account the source of the audio. In such a way that the audios emitted by the same person will only be in one of the partitions.

Finally, we carry out the training and testing of the network. We train using the same parameters as (Terbeh and Zrigui, 2017), changing the batch size to 4 and using 20 and 30 epochs.

3.2.3 Phonological Processes Identification

As can be seen in Fig. 5, the process to identify phonological processes is very similar to that described in Fig. 5. The only differences are the input and output data; that is, the AlexNet in charge of this process is trained only with the mispronounced words of the corpus and its completely connected layers are modified based on the tree classes of PP's (Structure of the syllable and word, Assimilation and Substitution).

4 EXPERIMENTS

In this section we will explain the experiments we carried out to see the performance of our solution.

4.1 Experimental Protocol

All of the experiments were performed using a Google Collaborate notebook. The free tier gives us a

Table 4: Results obtained by the models.

(a) Mispronunciation detection models.

	Accuracy	Precision	Recall	F-measure
Baseline (Franciscatto et al., 2021)	92.5%	-	-	-
Our Proposal (KS-20 epochs)	85.3%	83.1%	94.8%	88.6%
Our Proposal (KS-30 epochs)	87.8%	89.9%	89.5%	89.7%
Our Proposal (US-20 epochs)	82.8%	87.0%	83.4%	85.2%
Our Proposal (US-30 epochs)	**93.4**%	**93.5**%	**95.4**%	**94.5**%

(b) Phonological processes identification models.

	Accuracy	Precision	Recall	F-measure
Baseline (Franciscatto et al., 2021)	94.6%	-	-	-
Our Proposal (KS-20 epochs)	97.6%	98.3%	93.6%	95.8%
Our Proposal (KS-30 epochs)	**99.1**%	**98.7**%	**98.1**%	**98.4**%
Our Proposal (US-20 epochs)	89.6%	87.6%	78.5%	82.8%
Our Proposal (US-30 epochs)	94.7%	95.6%	86.7%	90.6%

(c) Comparison of trained models with and without voice conversion

	Accuracy	Precision	Recall	F-measure
Model trained with adult voices	77.0%	74.5%	**93.1**%	82.9%
Model trained with generated voices	**79.8**%	**82.2**%	84.0%	**83.1**%

virtual machine with a single-core Intel Xeon, 13GB of Ram and a 32GB HDD.

We used the programming language Python v.3.6. As well as the librosa v.0.8.1, scikit-image v.0.16.2 and PyTorch v.1.9.0+cu111 libraries.

Our dataset and code are freely available at https://github.com/Senzouz/WawaSimi

4.2 Results

As we used a different methodology than in (Franciscatto et al., 2021), we compared our Deep Learning proposal with the Machine Learning proposal presented in (Franciscatto et al., 2021). We trained our model using the combination of 20 and 30 epochs with both data-split approaches, Known Speaker and Unknown Speaker. The results of our experiments in terms of the mispronunciation detection model can be seen in Table 4.

In the case of the Phonological Processes Identification proposal, we performed the same experiments as in the Mispronunciation Detection method and compared it with the highest accuracy model reported in (Franciscatto et al., 2021). This results are shown in Table 4b.

Inspired by (Shahnawazuddin et al., 2020), we decided to replicate their experiments but with our methods to validate the effectiveness of the voice conversion. This means, we train a model with the 554 original audios (adult voices) and another model with the same amount of generated audios, which were randomly selected. Both models were tested with 222 generated audios selected for validation, the results obtained can be viewed in Table 4c.

4.3 Discussion

As can be seen in Table 4, our best model obtained 93.4% accuracy when detecting mispronunciation. This is probably because the model was trained with the Unknown Speaker, which is better suited for new instances than Known Speaker as talkers are only part of the training or testing phase, instead of being randomly distributed between both phases.

On the other hand, as shown in Table 4b, our best model obtained 99.1% accuracy when identify-

ing PP's. This may happen because in this process the acoustic characteristics of a voice are more influential. Therefore, the Known Speaker approach obtains better results because the model trains with all the voices available in the dataset learning their acoustic characteristics.

As can be seen in the Table 4c, the voice conversion method improves the performance of models when detecting mispronunciation. When we trained the model with only 554 synthetic audios the accuracy metric increased 2.8%. However, when using the full set of synthetic audios as in the Table 4, the performance increased by 26.4%.

5 CONCLUSIONS AND PERSPECTIVES

Thanks to the usage of a transfer-learnt AlexNet, we were able to develop a method to detect mispronunciation in Spanish as well as identify phonological processes with 93.4% and 99.1% respectively.

While language-learning apps like Duolingo, Babbel also employ classification methods to determine whether a word is partially or fully mispronounced (de la Cal Rioja, 2016). We carry out a second classification to determine the type of phonological process present in the misspronunciation.

Furthermore, machine and deep learning techniques are being used for various voice recognition tasks (e.g., query by humming (Alfaro-Paredes et al., 2021)), that could lead us to improvements in our approach. Additionally, text formalization (de Rivero et al., 2021) could be applied for educational purposes in compliment with our proposal.

For future works, we would like to have the support of a speech and language therapist who will help us with the construction of a data set that includes as many possible mispronunciation scenarios. In this way, we reduce the probability that the models do not know how to classify a new audio instance.

REFERENCES

Alfaro-Paredes, E., Alfaro-Carrasco, L., and Ugarte, W. (2021). Query by humming for song identification using voice isolation. In *IEA/AIE*.

Cappellari, V. M. and Cielo, C. A. (2008). Vocal acoustic characteristics in pre-school aged children. *Brazilian Journal of Otorhinolaryngology*, 74(2).

Charney, S. A., Camarata, S. M., and Chern, A. (2021). Potential impact of the covid-19 pandemic on communication and language skills in children. *Otolaryngology–Head and Neck Surgery*, 165(1).

Coloma, C. J., Pavez, M. M., Maggiolo, M., and Peñaloza, C. (2010). Desarrollo fonológico en niños de 3 y 4 años según la fonología natural: Incidencia de la edad y del género. *Revista signos*, 43(72).

de la Cal Rioja, J. (2016). Hound word. software para la mejora de la pronunciación en inglés. Universidad de Valladolid - Undergraduate thesis. https://uvadoc.uva.es/handle/10324/17963.

de Rivero, M., Tirado, C., and Ugarte, W. (2021). Formalstyler: GPT based model for formal style transfer based on formality and meaning preservation. In *KDIR*.

Franciscatto, M. H., Fabro, M. D. D., Lima, J. C. D., Trois, C., Moro, A., Maran, V., and Soares, M. K. (2021). Towards a speech therapy support system based on phonological processes early detection. *Comput. Speech Lang.*, 65.

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.*, 53(8).

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., and Zhang, G. (2015). Transfer learning using computational intelligence: A survey. *Knowl. Based Syst.*, 80.

Nazir, F., Majeed, M. N., Ghazanfar, M. A., and Maqsood, M. (2019). Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for arabic phonemes. *IEEE Access*, 7.

Pavez, M. M. and Coloma, C. J. (2017). Phonological problems in spanish-speaking children. *Advances in Speech-language Pathology*.

Popa, V., Silén, H., Nurminen, J., and Gabbouj, M. (2012). Local linear transformation for voice conversion. In *ICASSP*. IEEE.

Schuckman, M. (2008). *Voice characteristics of preschool age children*. PhD thesis, Miami University.

Shahnawazuddin, S., Adiga, N., Kathania, H. K., and Sai, B. T. (2020). Creating speaker independent ASR system through prosody modification based data augmentation. *Pattern Recognit. Lett.*, 131.

Terbeh, N. and Zrigui, M. (2017). Identification of pronunciation defects in spoken arabic language. In *PACLING*, volume 781 of *Communications in Computer and Information Science*.

Wu, Z. and Li, H. (2014). Voice conversion versus speaker verification: an overview. *APSIPA Transactions on Signal and Information Processing*, 3.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 8(4).