INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Measuring Phonological Precision in Children with Cleft Lip and Palate

*T. Arias-Vergara*[1,2,3], *E. Londoño-Mora*[2], *P. A. Pérez-Toro*[1,2], *M. Schuster*[3], *E. Nöth*[1], *J. R. Orozco-Arroyave*[1,2], *A. Maier*[1]

[1]Pattern Recognition Lab. Friedrich-Alexander Universität, Erlangen-Nürnberg, Erlangen, Germany
[2]GITA Lab, Facultad de Ingeniería. Universidad de Antioquia (UdeA), Medellín, Colombia
[3]Department of Otorhinolaryngology, Head and Neck Surgery. Ludwig-Maximilians University, Munich, Germany

corresponding author: tomas.arias@fau.de

## Abstract

Children with Cleft Lip and Palate (CLP) may experience difficulties in oral communication, leading to other developmental problems such as delayed language acquisition and poor social skills; thus, early treatment is essential for successful speech rehabilitation. In this paper, we propose a methodology for automatically assessing the phonological precision of children with CLP. We propose to use the probabilities obtained from a phonological class recognizer to measure phonological precision during connected speech. Furthermore, we compute the nasal-to-sound ratio to improve the automatic detection of the nasality level. For this, we considered speech recordings of 88 children with CLP, assessed by a clinician according to four nasality levels: *normal*, *mild*, *moderate*, and *severe*. We obtained an F1-score of up to 0.54 for detecting the nasality level automatically. The results suggest that phonological analysis can be used for individualized speech rehabilitation.

**Index Terms**: phonological analysis, deep learning, cleft lip and palate, pathological speech processing, children's speech.

## 1. Introduction

Oral communication of children with Cleft Lip and Palate (CLP) can be affected in several ways. CLP can cause Velopharyngeal Dysfunction (VPD), resulting in hypernasality, characterized by excessive resonance in the nasal cavity when producing vowels or voiced consonants [1]. VPD can also cause a significant nasal emission due to a large velopharyngeal opening, resulting in weak consonant production, short utterance length, and the development of compensatory articulation productions [2]. Clinical assessment is necessary to determine whether the child requires speech therapy, further surgery, or both. However, speech production can be impaired even after surgery [3]. Children with CLP should receive early and constant speech therapy to develop complete control of the pharyngeal velum in the shortest possible time. One possible way to increase the speed of rehabilitation is to implement assistive speech technology based on phonological analysis. By breaking down spoken words into their sound units, speech therapists can help children learn to produce and distinguish between different sounds more accurately. This process can involve teaching specific articulatory movements, such as how to position the tongue and lips for certain sounds. Through targeted phonological analysis and therapy, children with CLP can improve their speech intelligibility and overall communication [4].

### 1.1. Related work

Many studies in the literature have considered spectral-based analysis for automatic detection of speech nasalization in children with CLP [5, 6, 7, 8]. However, a relatively lower number of studies have considered phonological-based analysis for assisted assessment of speech nasalization [9]. The authors of [10] proposed automatic detection of articulation disorders in 58 German children with CLP, using several classifiers (late fusion) and an Automatic Speech Recognition (ASR) system to compute phoneme and word-level features such as Goodness of Pronunciation (GoP), phoneme posteriors, and word confidence score. Expert clinicians annotated the number of hypernasalized vowels and consonants, weakened pressure consonants, glottal articulation, and pharyngealizations. The authors obtained a Pearson's correlation coefficient of up to 0.89 between the percentage of mispronounced words and the rate of detected articulation errors. In [11], the authors considered speech recordings of 60 children (native American English speakers) with cleft palate to investigate the impact of forced-alignment errors in stop, fricative, nasal, approximant, and vowel sounds. For this, the authors computed the GoP using manual and automatic phoneme alignments and reported that nasal and vowel sounds impacted forced alignment the most. In [12], the authors proposed nasal distinctiveness features to assess hypernasal speech. Although the authors considered recordings from 75 speakers diagnosed with clinical conditions different than CLP, their method allows measuring nasalization in stop sounds (/p/, /t/, /b/, and /d/) by computing the log-likelihood ratio of the phoneme posterior probability and its corresponding nasal cognate, i.e., the nasal sound with a similar place of articulation. The authors performed regression analysis with the proposed features and found significant correlations with perceptual nasality scores.

### 1.2. Contributions of this work

In this paper, we propose an automatic method for measuring the **phonological precision** of children with CLP. An expert clinician assessed the nasalization level of the children according to four levels: *normal*, *mild*, *moderate*, and *severe*. Phonological precision is measured using **posterior probabilities** computed from the output of a recurrent neural network with Long-Short Term Memory cells (LSTM) and convolutional layers trained to automatically detect speech sounds according to voicing, manner, and place of articulation. We called this system ***PhonoQ*** and is based on the work presented in [13]. The probabilities computed with the network are used to predict the phonological class of a speech segment. We assume that these probabilities quantify "how well a sound was understood" by *PhonoQ*; thus, it measures the phonological precision. For comparison, we also estimate the time alignments using a semi-supervised approach: forced alignment is performed with a web-based ASR, and the phonological precision is measured with *PhonoQ*. We also compute the **nasal-to-sound ratio** to measure the amount of nasalization in a speech segment.


10.21437/Interspeech.2023-2099

## 2. Materials and Methods

### 2.1. Data

Speech recordings of 88 children with CLP were considered for analysis. All children are native Spanish speakers from Colombia. The recordings were sampled at 16kHz with a 16-bit resolution. An expert clinician assessed the degree of speech nasalization of the children according to four levels: *normal* (20 children), *mild* (28 children), *moderate* (25 children), and *severe* (15 children). The children were asked to read the sentences reported in Table 1

Table 1: *Sentences read by the children with CLP.*

| Key | Sentence | IPA translation |
|---|---|---|
| Carlos | Carlos coge su pelota | karlos koxe su pelota |
| Gato | El gato toma leche | el gato toma letʃe |
| Llave | La llave de la casa | la ʎaβe de la kasa |
| Silla | La silla es cafe | la siʎa es kafe |
| Susi | Susi come sopa | susi kome sopa |
| Tomas | Tomas toca tambor | tomas toka tambor |

### 2.2. Phonological analysis

#### 2.2.1. Phonological class recognition

We implemented an automatic phonological recognizer (*PhonoQ*[1]) using an LSTM to predict 18 phonological classes (including silence) according to voicing, manner, and place of articulation (see Table 2). Furthermore, we trained the LSTM

Table 2: *Phonological classes considered in this study.*

| | Dimension | Class | Examples (IPA) |
|---|---|---|---|
| 0 | - | Silence | - |
| 1 | Manner | Stop | /p/, /t/, /k/, /b/, /d/, /g/ |
| 2 | | Nasal | /n/, /m/, /ŋ/ |
| 3 | | Trill | /r/, /ɾ/ |
| 4 | | Fricative | /s/, /ʃ/, /z/, /f/ |
| 5 | | Approximants | /j/ |
| 6 | | Lateral | /l/ |
| 7 | | Vowel | /a/, /e/, /i/, /o/, /u/ |
| 8 | Place | Labial | /p/, /b/, /m/, /f/, /v/ |
| 9 | | Alveolar | /t/, /d/, /n/, /ɾ/, /s/, /z/, /l/ |
| 10 | | Velar | /k/, /g/, x, /ŋ/ |
| 11 | | Palatal | /j/, /ʎ/ |
| 12 | | Postalveolar | /ʃ/ |
| 13 | | Central | /a/, /aː/ |
| 14 | | Front | /i/, /e/ |
| 15 | | Back | /u/, /o/ |
| 16 | Voicing | Voiceless | /p/, /t/, /k/, /ʃ/, /s/ |
| 17 | | Voiced | /m/, /n/, /b/, /d/, /g/, /a/ |

as a **multilabel** classifier because some speech sounds can be part of multiple classes, e.g., /p/ belongs to the "stop", "labial", and "voiceless" classes. The network predicts phonological sequences by computing the probability of the occurrence of a phonological class in a speech recording; thus, the output consists of sequences of **phonological posterior probabilities**. Figure 1 shows an example of the sequence of phonological posterior probabilities computed from the recording for the Spanish word "sopa" (IPA: /sopa/).

[1] https://github.com/TAriasVergara/PhonoQ

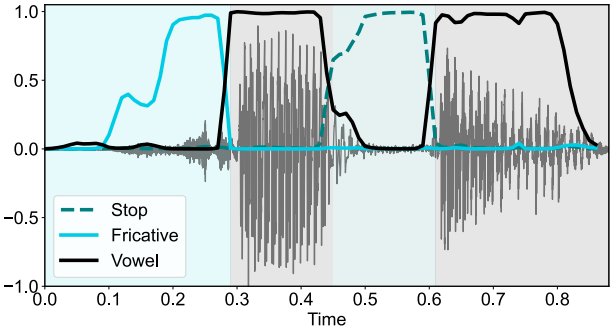


Figure 1: *Sequence of phonological posterior probabilities computed from the recording of the Spanish word "sopa".*

#### 2.2.2. Training

The architecture of *PhonoQ* is shown in Figure 2. We trained the model with the TEDx Spanish Corpus (TSC)[2], a dataset containing 24 hours and 29 minutes of speech recordings (from TEDx events) from 142 Spanish native speakers. The input tensors are chunks of 500 ms extracted from the speech signals. Then, Mel-spectrograms are computed for each chunk using 64 triangular filters and Hanning windows of 25 ms with a step size of 10 ms. The convolutions are performed with a 3×3 kernel and ReLU activation functions; however, we use a padding of 1 in the time axis to keep a one-to-one relation between the length of the input (speech sequence) and the output (phonological prediction). Max-pooling is performed only in the frequency axis with a 1×2 kernel. We optimize the model using the Adam algorithm with a learning rate of $\eta = 10^{-4}$ and binary cross-entropy logistic loss. Furthermore, we used class weights (computed in the training set) in the loss function to account for class unbalance, i.e., some phonological classes are underrepresented. We use a batch size of 100 and consider an *early stopping* strategy of seven epochs. The output of the phonological class recognizer is the posterior probabilities computed with a sigmoid activation function. To predict the phonological sequence from a recording is necessary to divide the signal into chunks of 500 ms; thus, we align the phonological predictions by concatenating the output of the LSTM.

#### 2.2.3. Unsupervised measure of phonological precision

The phonological class recognizer makes predictions based on posterior probabilities. In this study, we use these probabilities to measure phonological precision in children with CLP and investigate the influence of nasalization on speech production. We assume that the phonological class recognizer is trained with "good spoken" Spanish; thus, the posterior probability indicates how well the system "understands" certain sounds, i.e., the closer the phonological class probability to "1", the better a speaker pronounces it. Note that in Figure 1, the posterior probabilities decrease in the transition from one predicted phonological class to another, which is normal because it reflects the movement of the articulators when going from one sound to another. For this reason, we measure phonological precision by computing the maximum posterior probability (MaxPh) of the predicted phonological segment. Additionally, since a speech recording may contain several speech segments for the same phonological class, we compute the average MaxPh per phonological class. Figure 3 shows the radar plot of the average MaxPh computed for children with *normal*, *mild*, *moderate*,

[2] http://www.ciempiess.org/downloads

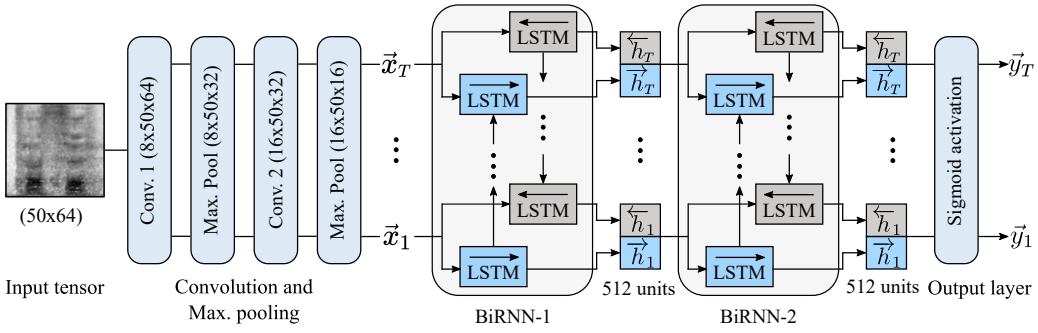


Figure 2: *Architecture of PhonoQ.*

and *severe* nasalization levels. The figure shows that "stop" and

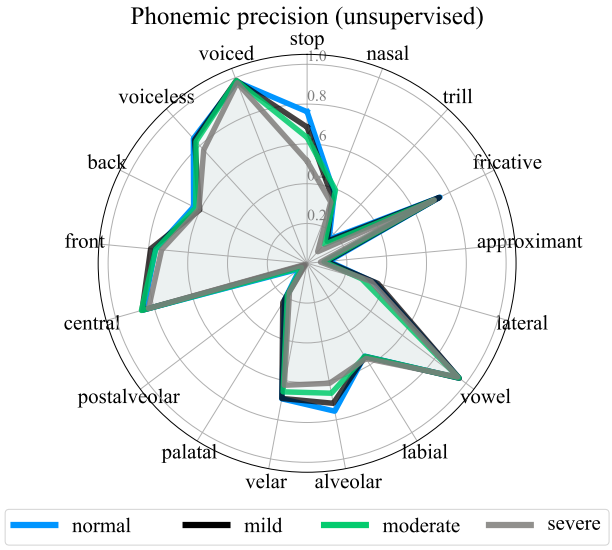


Figure 3: *Phonological precision (average* MaxPh*) of the CLP children using the unsupervised approach.*

"alveolar" phonological classes have the highest difference between groups, "back" vowels have the lowest phonological precision compared to "central" and "front" vowels, and in general, children with *severe* nasalization have lower MaxPh compared to the other groups. Note also that the phonological precision for "vowel" is relatively high, even though the "back" class have relatively low values. The reason is that "vowel" contains all vowels regardless of the place of articulation; thus, the general precision is higher, which shows the importance of considering different phonological categories for speech assessment.

#### 2.2.4. *Nasal-to-sound ratio*

Additional to the phonological precision, we also computed the nasal-to-sound ratio, which in this study is defined as the level of nasalization in a phonological class. We compute this measure as

$$\mathrm{rN} = \frac{P(n|p)}{\mathrm{MaxPh}_p} \tag{1}$$

where $\mathrm{MaxPh}_p$ is the (maximum) probability that a speech segment is a phonological class $p$ and $P(n|p)$, is the probability that $p$ is a nasal sound. Values of rN close to 1 (or higher) represent high nasalization.

#### 2.2.5. *Semi-supervised measure of phonological precision*

We performed forced alignment on the recording to measure phonological precision using the posterior probabilities computed from our recognizer. We used the BAS CLARIN web service [14] to get the time stamps of every phonological class in the recording and labeled them according to the classes reported in Table 2. Then, we computed MaxPh and rN for every phonological class in the intervals obtained with the ASR. Figure 4 shows the phonological precision measured with the semi-supervised approach. Compared to the unsupervised approach, the difference between children with *normal*, *mild*, *moderate*, and *severe* nasalization is more evident for the "stop", "alveolar", "velar", and "labial" phonological class.

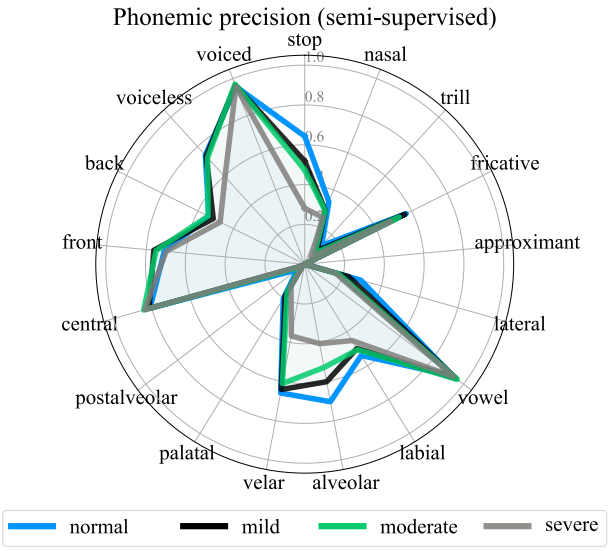


Figure 4: *Phonological precision (average* MaxPh*) of the CLP children using the semi-supervised approach.*

### 2.3. Automatic classification

We trained a linear multiclass Support Vector Machine (SVM) to predict the nasality level of the children. The margin parameter $C$ was optimized through a grid search with $2^{-7} < C < 2^3$ using a nested 10-fold cross-validation strategy, i.e., for every fold, an internal 9-fold cross-validation is performed to find the optimum $C$. We tested the performance of the SVM by computing the median of the best $C$ obtained during training; then, we performed cross-validation again with fixed parameters. The

classifier's performance is evaluated with the F1-score, the precision, and the recall. We used a linear SVM to avoid overfitting due to the limited sample size.

### 2.4. Feature importance

We ranked the feature intelligibility features using *permutation feature importance*, an inspection technique used to evaluate the dependence of a model (e.g., the SVM) on a set of features[3]. In this study, the importance of a feature is defined as the decrease in the F1-score when a single feature is randomly shuffled between samples, thus, breaking the relationship between feature and target. The feature importance was evaluated during training. Then, the average importance score is computed to obtain the final ranking of features.

## 3. Experiments and Results

We performed automatic classification of children with *normal*, *mild*, *moderate*, and *severe* nasalization levels using the phonological precision and nasal-to-sound ratio. This classification task is important to have an overall assessment (i.e., a second opinion for the clinician) of the progression of the speech rehabilitation, which can be complemented with individual analysis of phonological precision. The classification is performed for each sentence to investigate the dependency on the phonetic content for assessing nasality, e.g., the keyword sentences "Carlos", "Llave", and "Silla" contains stop, fricative, and lateral sounds but not nasals. Table 3 shows the classification results with and without feature importance analysis for the unsupervised and semi-supervised approaches. Overall, feature selection improved the automatic detection of nasality levels for all sentences. In the case of the supervised method, the best results were obtained with the sentences "Carlos" (F1=0.54) and "Susi" (F1=0.50), while for the semi-supervised the best results were obtained with "Silla" (F1=0.52) and "Tomas" (F1=0.50). To better understand this mismatch, we plotted the word clouds of the top 10 ranked features (per sentence) for the unsupervised and semi-supervised approaches. Figure 5 shows the obtained results. In the figure, "alveolar" (e.g., /t/, /d/, /n/) and "voiceless" sounds (/t/, /p/, /s/) appear to be among most of the ranked features for both unsupervised and semi-supervised approaches. For the unsupervised approach, "labial", "lateral," and "central" phonological classes were more frequently ranked as important across sentences. For the semi-supervised approach, the most important features were the "velar", "stop," and "vowel" sounds. Nasal-to-sound ratios were also among the top 10 features across sentences; however, their ranking was lower than the phonological precision features.

Table 3: *Results for automatic detection of nasality levels using unsupervised and semi-supervised approaches.* ***Task:*** *sentence keyword.* ***# feats:*** *number of features.* ***Prec:*** *precision.* ***Rec:*** *recall.* ***F1:*** *F1-score*

| Unsupervised | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Task** | **All features** | | | | **Feature importance** | | | |
| | **# feats** | **Prec** | **Rec** | **F1** | **# feats** | **Prec** | **Rec** | **F1** |
| Carlos | 34 | 0.46 | 0.44 | 0.45 | 11 | **0.59** | **0.54** | **0.54** |
| Gato | 34 | 0.30 | 0.29 | 0.29 | 6 | 0.43 | 0.44 | 0.43 |
| Llave | 34 | 0.28 | 0.27 | 0.26 | 12 | 0.42 | 0.38 | 0.38 |
| Silla | 34 | 0.24 | 0.24 | 0.24 | 18 | 0.36 | 0.34 | 0.35 |
| Susi | 34 | 0.31 | 0.31 | 0.30 | 2 | **0.56** | **0.51** | **0.50** |
| Tomas | 34 | 0.30 | 0.31 | 0.30 | 14 | 0.36 | 0.36 | 0.35 |
| Average | - | 0.31 | 0.31 | 0.31 | - | 0.45 | 0.43 | 0.42 |
| **Semi-supervised** | | | | | | | | |
| **Task** | **All features** | | | | **Feature importance** | | | |
| | **# feats** | **Prec** | **Rec** | **F1** | **# feats** | **Prec** | **Rec** | **F1** |
| Carlos | 34 | 0.45 | 0.44 | 0.44 | 9 | 0.50 | 0.49 | 0.49 |
| Gato | 34 | 0.36 | 0.35 | 0.34 | 13 | 0.44 | 0.42 | 0.43 |
| Llave | 34 | 0.31 | 0.32 | 0.31 | 7 | 0.47 | 0.44 | 0.45 |
| Silla | 34 | 0.36 | 0.34 | 0.34 | 9 | **0.54** | **0.51** | **0.52** |
| Susi | 34 | 0.42 | 0.42 | 0.41 | 15 | 0.49 | 0.48 | 0.47 |
| Tomas | 34 | 0.36 | 0.32 | 0.33 | 12 | **0.54** | **0.48** | **0.50** |
| Average | - | 0.38 | 0.36 | 0.36 | - | 0.50 | 0.47 | 0.48 |

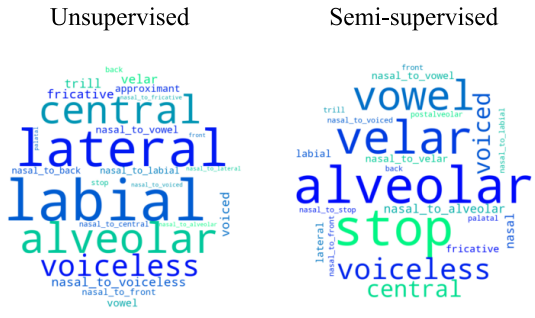


Figure 5: *Feature importance analysis for the unsupervised and supervised methods.*

## 4. Discussion and Conclusion

We presented a methodology to automatically estimate phonological precision from children with CLP on four levels of nasalization: *normal*, *mild*, *moderate*, and *severe*. We trained a phonological class recognizer called *PhonoQ* to measure phonological precision and the nasal-to-sound ratio. Unsupervised and semi-supervised approaches were used to obtain the time alignments of the phonological classes. When we compared the radar plots of the phonological precision obtained with the two approaches, we observed a similar trend on the affected sounds, i.e., the "back" vowels are the most affected by nasalization, and "alveolar" and "stop" sounds exhibits the higher difference between groups. However, these differences can be observed better in the radar plot obtained with forced alignment, which we expected due to the better match between the measured phonological precision in the actual time interval. Additionally, we performed a feature importance analysis for every sentence uttered by the children. As expected, feature selection improved the identification of the nasality level. When we compared the top 10 features from each sentence, we observed that "alveolar" sounds were ranked as the most important using the unsupervised and semi-supervised methods, i.e., "alveolar" sounds were the most affected by nasalization for our sample group. Our results are promising but not conclusive, as we have to consider the limitations of our dataset. For instance, with the available number of samples, it is not wise to generalize our findings regarding phonemic precision errors or feature importance selection; however, it is a good starting point considering the differences between groups.

## 5. Acknowledgements

The authors thank the Signal Processing and Recognition Group from the Universidad Nacional de Manizales for providing the data. This work received support from the Universidad de Antioquia through the project PI2023-58010.

[3] https://scikit-learn.org/stable/modules/permutation_importance.html

# 6. References


[1] R. Wyatt, D. Sell, J. Russell, A. Harding, K. Harland, and L. Albery, "Cleft palate speech dissected: a review of current knowledge and analysis," *British Journal of Plastic Surgery*, vol. 49, no. 3, pp. 143–149, 1996.

[2] A. W. Kummer, *Cleft palate & craniofacial anomalies: Effects on speech and resonance*. Nelson Education, 2013.

[3] A. C. Williams, D. Bearn, S. Mildinhall *et al.*, "Cleft lip and palate care in the United Kingdom—the Clinical Standards Advisory Group (CSAG) Study. Part 2: dentofacial outcomes and patient satisfaction," *The Cleft palate-craniofacial journal*, vol. 38, no. 1, pp. 24–29, 2001.

[4] K. J. Kotlarek and B. I. Krueger, "Treatment of Speech Sound Errors in Cleft Palate: A Tutorial for Speech-Language Pathology Assistants," *Language, Speech, and Hearing Services in Schools*, pp. 1–18, 2023.

[5] P. Vijayalakshmi, M. R. Reddy, and D. O'Shaughnessy, "Acoustic analysis and detection of hypernasality using a group delay function," *IEEE Transactions on biomedical engineering*, vol. 54, no. 4, pp. 621–629, 2007.

[6] T. Dodderi, M. Narra, S. M. Varghese *et al.*, "Spectral Analysis of Hypernasality in Cleft Palate Children: A Pre-Post Surgery Comparison," *Journal of clinical and diagnostic research: JCDR*, vol. 10, no. 1, p. MC01, 2016.

[7] M. Golabbakhsh, F. Abnavi, M. Kadkhodaei Elyaderani *et al.*, "Automatic identification of hypernasality in normal and cleft lip and palate patients with acoustic analysis of speech," *The Journal of the Acoustical Society of America*, vol. 141, no. 2, pp. 929–935, 2017.

[8] H. Carvajal-Castaño and J. R. Orozco-Arroyave, "Articulation analysis in the speech of children with cleft lip and palate," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings 24*. Springer, 2019, pp. 575–585.

[9] M. Shahin, U. Zafar, and B. Ahmed, "The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 400–412, 2019.

[10] A. Maier, F. Hönig, T. Bocklet *et al.*, "Automatic detection of articulation disorders in children with cleft lip and palate," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2589–2602, 2009.

[11] 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.

[12] M. Saxon, J. Liss, and V. Berisha, "Objective measures of plosive nasalization in hypernasal speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6520–6524.

[13] T. Arias-Vergara, *Analysis of Pathological Speech Signals*. Germany: Logos Verlag Berlin, 2022.

[14] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.