Interspeech 2018
2-6 September 2018, Hyderabad

# Detection of Glottal Activity Errors in Production of Stop Consonants in Children with Cleft Lip and Palate

*C M Vikram[1], S R Mahadeva Prasanna[1,2], Ajish K Abraham[3], Pushpavathi M[3], Girish K S[3]*

[1] Indian Institute of Technology Guwahati, Guwahati, India
[2]Indian Institute of Technology Dharwad, Dharwad, India
[3]All India Institute of Speech and Hearing, Mysuru, India

{cmvikram, prasanna}@iitg.ernet.in, ajish68@aiishmysore.in, shivanna.pushpa@gmail.com, girish06ks@gmail.com

## Abstract

Individuals with cleft lip and palate (CLP) alter the glottal activity characteristics during the production of stop consonants. The presence/absence of glottal vibrations during the production of unvoiced/voiced stops is referred as glottal activity error (GAE). In this work, acoustic-phonetic and production based knowledge of stop consonants are exploited to propose an algorithm for the automatic detection of GAE. The algorithm uses zero frequency filtered and band-pass (500-4000 Hz) filtered speech signals to identify the syllable nuclei positions, followed by the detection of glottal activity characteristics of consonant present within the syllable. Based on the identified glottal activity characteristics of consonant and *a priori* voicing information of target stop consonant, the presence or absence of GAE is detected. The algorithm is evaluated over the database containing the responses of normal children and children with repaired CLP for the target consonant-vowel-consonant-vowel words with stop consonants.

**Index Terms**: Articulation errors, cleft lip and palate, glottal activity errors, and stop consonants.

## 1. Introduction

Glottal activity refers to the quasi-periodic vibration of vocal folds during the production of voiced speech [1]. Glottal activity during the articulatory closure phase of stops is considered as a primary feature to discriminate the voiced and unvoiced stops. Articulatory closure phase of voiced stops is characterized by a weakly voiced signal (voice bar), whereas silence in case of unvoiced stops [2]. During the articulatory closure phase of stops, it is necessary to close both oral and nasal cavities, in-order to develop the intraoral air pressure ($P_0$). The speakers with cleft lip and palate (CLP) are unable to build adequate $P_0$, due to the loss of airflow through the velopharyngeal gap or oro-nasal fistula. As a result, CLP speaker greatly alters the characteristics of stops and produces articulation errors.

Different types of articulation errors such as nasalized stops, weak stops, devoicing errors, glottal, pharyngeal, palatal, and nasal substitutions are reported in the literature related to CLP speech [3, 4, 5, 6]. Among these errors, nasal substitution for unvoiced stops is produced with highly altered glottal activity characteristics. During the production of unvoiced stops, the loss of $P_0$ increases the pressure across the glottis. This may initiate the glottal vibrations during the articulatory closure period of unvoiced stops. Replacement of silence bar of unvoiced stop by voiced nasal consonants is reported in Ref. [7]. During the production of voiced stops, speakers with CLP may suppress the glottal vibrations by completely opening or closing of the vocal folds. This will result in the production of devoicing error or glottal stops. The absence of glottal vibrations during voiced stops and their presence during the production of unvoiced stops are collectively referred as glottal activity errors (GAEs).

Stops constitute a major class of sound units in a language. Hence, the assessment of stops is considered to be important during the diagnosis and therapy of CLP speakers. Objective evaluation of articulation errors using signal processing techniques will be helpful for the speech-language pathologists (SLPs) [8, 9]. In the literature, most of the works are focused on the detection of hypernasality in CLP speakers [10, 11, 12] . With respect to articulation errors, methods for the detection of laryngeal backing, pharyngeal backing, nasalized consonants, weak pressure consonants, and glottal stops are proposed [8, 13]. However, as per the knowledge of existing literature, no works have been addressed the detection of GAEs in CLP speakers. Detection of GAE gives the important information about the deviant glottal source mechanism during the production of stops. This excitation source specific information may help SLPs to correct the deviant glottal vibration mechanism in CLP speakers.

The current work is mainly motivated to develop a signal processing method for the detection of GAE during the production of stop consonants by speakers with CLP. The proposed GAE detection algorithm uses acoustic-phonetic features derived from zero frequency filtered and band-pass filtered speech signals. The paper is organized as follows: Section 2 describes the database and perceptual evaluation. Algorithm for the automatic detection of GAE is presented in Section 3. The experimental results are discussed in Section 4. The conclusion and the possible future directions are mentioned in Section 5.

## 2. Database and Perceptual Evaluation

In this work, 37 children with repaired CLP in the age range of 6 to 12 years were considered. 30 typically developed children were considered in the same age range. All the participants considered for the study had Kannada as their native language. The language abilities of all the participants with repaired CLP were age adequate. Individuals with other associated problems like hearing loss, intellectual disability, and nasal pathologies were excluded from this study.

Speech stimuli are comprised of 8 non-meaningful disyllabic consonant-vowel-consonant-vowel (CVCV) words containing stop consonants. Unvoiced stop consonants: bilabial /p/, dental /t/, retroflex /T/, and velar /k/; their corresponding voiced cognates: /b/, /d/, /D/ and /g/ are used with the combination of vowel /a/ to form the CVCV words like /papa/, /tata/, /TaTa/, etc. Participants were asked to repeat the words after the tester. The responses were recorded in a sound-treated room with a


10.21437/Interspeech.2018-1665

sampling frequency of 48,000 Hz and digitized at 16 bits per samples using Bruel & Kjaer sound level meter (type 2250-s handheld analyzer). All the recorded samples are analyzed us-

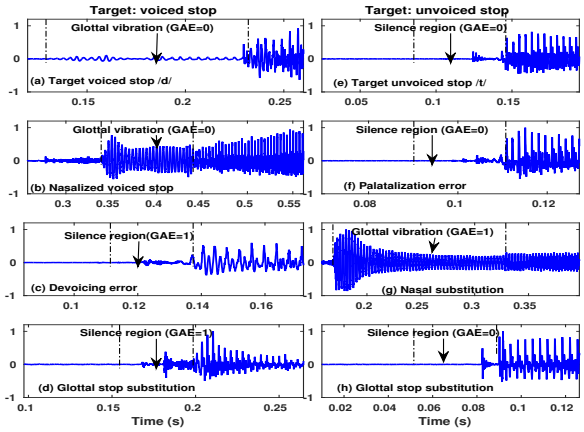


Figure 1: *Illustration of waveform based cues for the analysis of glottal activity error (GAE). (a) normal voiced stop (/d/), (b)-(d) nasalized voiced stop, devoicing error, and glottal stop substitution produced for target voiced stop (/d/). (e) normal unvoiced stop (/t/), (f)-(h) palatalization error, nasal, and glottal stop substitution for the target unvoiced stop /t/.*

ing PRAAT software [14]. The vowels and consonant regions are manually marked. Further, based on the presence of periodic voiced signal in the speech waveform and voicing information of the target stop, the presence or absence of GAE is annotated. The presence of GAE is denoted as GAE=1, whereas its absence by GAE=0. The conventions used for the annotation of GAE by signal based evidence are illustrated in Figure 1. The target voiced stop depicted in Figure 1(a) shows the presence of glottal vibration. The absence of glottal vibrations for the target voiced stop, i.e., devoicing error and glottal stops (Figures 1(c) and (d)) are considered as the cases with the presence of GAE (GAE=1). For the target voiced stop, nasalized voiced stop (Figures 1(b)) shows the presence of glottal vibrations, which is marked as GAE=0. For the target unvoiced stops (Figure 1(g)), the voiced substitutions such as nasal consonant are marked as GAE=1. For the unvoiced stops, absence of glottal vibrations as in the case of palatalized and glottal stop (Figures 1(f) and (h)) are marked as GAE=0. The total number of syllables containing with GAE=0 and GAE=1 for normal and CLP speakers are presented in Table 1.

In order to compare the signal and perceptual evaluation based identification of GAE, the recorded speech samples are presented to three experienced SLPs. Each SLP is asked to write the phonetic transcription of the word. The transcribed consonants are then mapped into consonants with GAE=1 and GAE=0 as follows. For the voiced target stop, if the transcription contains any unvoiced sound or complete glottal stop, then it is considered as GAE=1. Whereas, if the transcription contains voiced consonant then it is treated as GAE=0. Similarly, for target unvoiced stop, the presence of any voiced substitutions such as nasal sounds, semivowels are grouped into GAE=1. Whereas for target unvoiced stop, any unvoiced substitutions, and complete glottal stops are grouped into GAE=0.

The reliability of GAE identification using signal based evidence (SBE) and perceptual evaluation is validated using Cohen's Kappa measure. Cohen's Kappa coefficient between $1^{st}$ SLP and SBE, $2^{nd}$ SLP and SBE, and $3^{rd}$ SLP and SBE are found to be 0.96, 0.95, and 0.99, respectively. Cohen's Kappa measure indicates that there is a significantly good agreement found between SBE and SLP's evaluation.

Table 1: *Description of database with the number of stop consonants with glottal activity error absent: GAE (0) and present: GAE (1)*

| Group | No. of Speakers | Target: voiced stop | | Target: unvoiced stop | |
|---|---|---|---|---|---|
| | | GAE=0 | GAE=1 | GAE=0 | GAE=1 |
| Normal | 25 (13 Male+ 12 Female) | 2400 | 0 | 2500 | 0 |
| CLP | 37 (21 Male + 16 Female) | 1500 | 850 | 2100 | 110 |

## 3. Method for The Detection of Glottal Activity Errors

The proposed algorithm for the detection of GAE involves different stages such as the detection of glottal activity regions, syllable nuclei locations, and voicing nature of consonant present in the syllable.

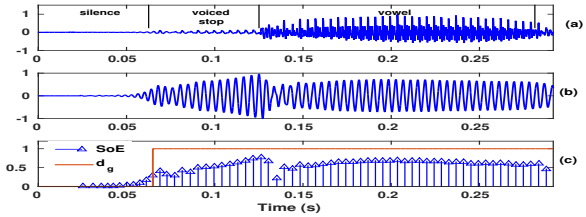


Figure 2: *Detection of glottal activity regions. (a) Speech waveform of syllable /da/ produced by normal speaker, (b) ZFFS, and (c) SoE superimposed with glottal activity decision ($d_g$).*

### 3.1. Detection of glottal activity regions

Zero frequency filter based approach is used for the detection of glottal activity regions. Zero frequency filtering of speech signal consists of the passage of differenced speech signal through a cascade of two ideal zero Hz resonators [15]. The output of the resonator contains cumulative DC bias, which is removed by local mean subtraction process. The local mean subtracted signal is termed as zero frequency filtered signal (ZFFS). The positive or negative going zero crossings of ZFFS correspond to the glottal closure instants or epoch locations. The first order slope of ZFFS, computed at epoch location is referred as the strength of excitation (SoE). Figures 2(a)-(c) represent the speech signal, ZFFS, and SoE computed at each epoch location, respectively. The SoE is relatively higher at voiced regions when compared to unvoiced or silence regions. Because, ZFFS captures the strength of quasi-periodic impulse-like excitation caused by glottal vibrations present around the zero frequency. Whereas the absence of glottal vibrations in silence or unvoiced regions shows the lower values of SoE.

If the SoE computed at each epoch locations is represented by $\gamma[n]$, then the binary evidence for the presence/absence of glottal activity $d_g[n]$ is computed as,

$$d_g[n] = \begin{cases} 1 \; if \; \gamma[n] > T_1, \\ 0 \; \; otherwise. \end{cases} \quad (1)$$

where the threshold $T_1$ is chosen equal to the 0.3 times of mean SoE, computed over entire utterance. Figure 2(c) shows the $d_g[n]$ superimposed over the SoE evidence. ZFFS captures the glottal activity associated with weakly voiced voice bar regions also (Figure 2(c)).

### 3.2. Syllable nuclei detection

In this work, the recorded words are disyllabic in nature, where each syllable consists of a high energy vowel and a consonant. Generally, in a CV syllable, the energy of vowels is relatively higher than that of consonants. However, in CLP speakers the energy contrast between vowels and consonants is significantly reduced due to the nasalization effect. Most of the energy of nasalized components is concentrated at lower frequencies, i.e., around the first formant of nasal consonants (approximately 500 Hz). Also, the most of the energy of vowels is concentrated below 4 kHz. Therefore, a bandpass filter with passband frequencies from 0.5 to 4 kHz is used to enhance the contrast between vowels and consonants. Figure 3(a) shows the speech waveform of CVCV word containing nasalized voiced stops, and its spectrogram is depicted in Figure 3(b). The bandpass filtered speech (BPFS) is shown in Figure 3(c). When compared to speech signal (Figure 3(a)), BPFS (Figure 3(c)) shows enhanced contrast between consonants and vowels.

The epoch synchronous short-term energy of BPFS is computed using windowed frames of 20 ms anchored around each epoch. The short-term energy contour is smoothed using 100 ms hamming window (Figure 3(d)). Within the glottal activity regions, the peaks of smoothed BPFS energy profile are detected to locate the syllable nuclei. The detected syllable nuclei are shown in Figure 3(d).

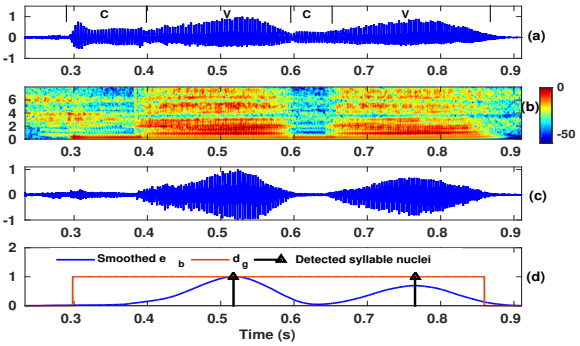


Figure 3: *Identification of syllable nuclei. (a) Speech signal of CVCV word containing nasalized voiced stop, (b) wide band spectrogram, (c) band pass (0.5-4 kHz) filtered speech (BPFS), and (d) smoothed short term energy of BPFS ($e_b$), glottal activity region ($d_g$), and detected syllable nuclei.*

### 3.3. Detection of voiced consonants

Once the positions of syllable nuclei are determined, the voicing nature of consonant present within the syllable needs to be analyzed. In order to identify the voiced consonants with low-frequency dominant spectral characteristics, the ratio of ZFFS to BPFS is computed as $r_{zb}[n] = e_z[n]/e_b[n]$, where $e_z[n]$ and $e_b[n]$ are the epoch synchronous short-term energies of ZFFS and BPFS, respectively. Here, short-term energies are computed using 20 ms windowed frames anchored around epochs. Before computing the energy, ZFFS and BPFS are $l_2$ normalized. Figure 4(a) shows the speech waveform of CVCV word containing voiced consonant (nasalized voiced stop). Figure 4(b) shows the short-term energies of ZFFS and BPFS. In Figure 4(b), ZFFS shows relatively higher energy than BPFS at consonant regions. This is due to the fact that zero frequency filter acts as a bandpass filter around zero frequency, which allows most of the signal energy present around zero Hz. Therefore, zero frequency filter passes most of the energy present in the voiced consonants, while allowing only a part of the energy of vowels. The high-frequency energy of vowels is more than that of voiced consonants. So the energy of BPFS is higher at the vowels than that of consonant regions. As a result, the evidence $r_{zb}$ shown in Figure 4(c) indicates higher values at voiced consonants, when compared to vowel regions. Using $r_{zb}$ and glottal activity decision $d_g$, binary evidence $d_{lfdvr}$ for the detection of low-frequency dominant voiced regions (LFDVR) is computed as

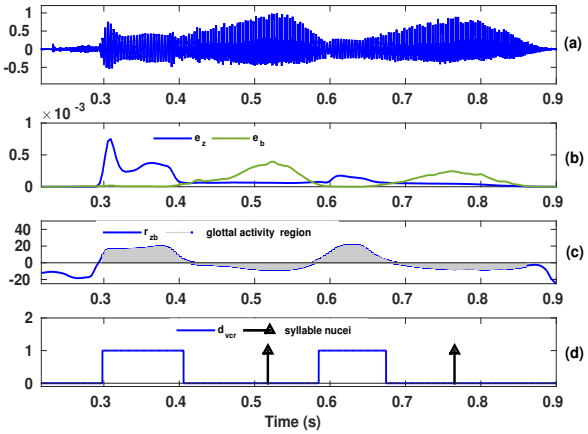


Figure 4: *Detection of voiced consonant regions. (a) Speech signal containing nasalized voiced stop consonant, (b) short-term energies of ZFFS ($e_z$) and BPFS ($e_b$), (c) ratio of $e_z$ to $e_b$ in dB, and (d) detected voiced consonant regions.*

$$d_{lfdvr}[n] = \begin{cases} 1 \; if \; r_{zb}[n] > T_2 \; \& \; d_g[n] = 1, \\ 0 \; otherwise. \end{cases} \quad (2)$$

where the threshold $T_2$ is given by,

$$T_2 = \frac{1}{M} \sum_{j=1}^{M} r_{zb}(v_j) + \beta \quad (3)$$

where $v_j$ is the location of $j^{th}$ syllable nucleus, $M$ is the number of detected syllable nuclei, and $\beta$ is the relative difference of $r_{zb}$ between vowels and low frequency dominant voiced consonants. $\beta$ is estimated using a development set, comprised of 50 CVCV words containing voiced consonants. For each word, the difference between average values of $e_{zb}$ is measured at the manually marked vowel and voiced consonant regions. The mean and standard deviation of difference values across 50 words is found to be 20$\pm$10 dB. The lower bound of the distribution is found 10 dB, which is used as the $\beta$ value to segment the LFDVRs. The decision curve $d_{lfdvr}$, by indicating detected voiced consonant regions using LFDVR evidence is depicted in Figure 4(d). The detected LFDVRs below the minimum duration of a phoneme (30 ms) are considered as spurious regions and removed from the further analysis. Within the search interval $t_j$ defined around the $j^{th}$ syllable nucleus $v_j$, the consonant associated with the $j^{th}$ syllable is characterized as voiced or unvoiced. The voicing decision for the consonant present in $j^{th}$ syllable is given by

$$d_{vc_j} = \begin{cases} 1 \; if \; r_{zb}[\tau] = 1, \tau \in t_j, \\ 0 \; otherwise. \end{cases} \quad (4)$$

For word initial syllable $v_1$, the search interval $t_1$ is chosen from the beginning of the utterance to the location of syllable nucleus

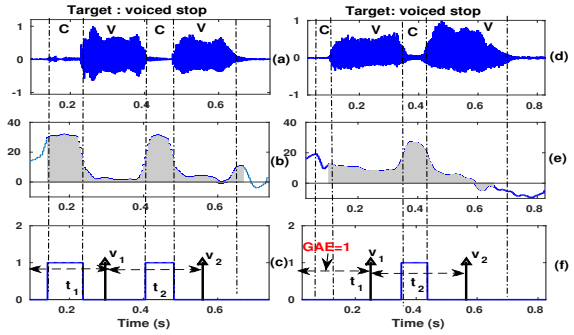


Figure 5: *(a)-(c) and (d)-(f) represent the speech waveform, ratio of ZFF to BPF ($r_{zb}$) in dB, and low frequency dominant voiced consonant region evidence ($d_{lfdvr}$) for the response of normal and CLP speakers for target word /dada/, respectively. $t_1$ and $t_2$ are the search intervals associated with syllable nuclei $v_1$ and $v_2$, respectively for the detection of voiced consonants. Evidence $d_{lfdvr}$ in subplot (f) indicates the absence of glottal vibrations for the target voiced consonant associated with first syllable.*

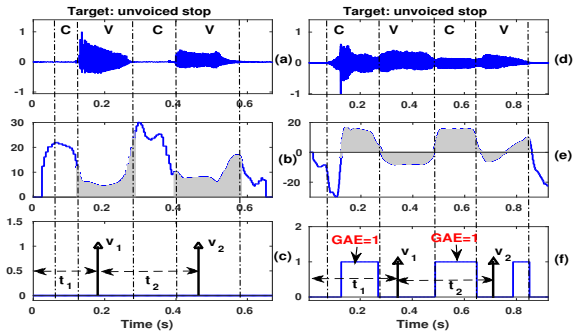


Figure 6: *(a)-(c) and (d)-(f) represent the speech waveform, ratio of ZFF to BPF ($r_{zb}$) in dB, and low frequency dominant voiced region evidence ($d_{lfdvr}$) for the response of normal and CLP speakers for target word [papa], respectively. $t_1$ and $t_2$ are the search intervals associated with syllable nuclei $v_1$ and $v_2$, respectively for the detection of voiced consonants. In subfigure (f), nasal substitution for unvoiced stop indicates the presence of GAE (GAE=1).*

$v_1$, i.e., $t_j \in [0, v_j], j = 1$ . Whereas for word medial syllable, $t_j$ is chosen as the interval between previous and current syllable nuclei i.e., $t_j \in [v_{j-1}, v_j], j \neq 1$.

### 3.4. Decision of GAE

Based on the presence of voiced consonant and *a prior* voicing information of target stops, the GAE is determined as

$$\widehat{GAE}_j = \begin{cases} 1 \; if \; Target = voiced \; stop \; \& \; d_{vc} = 0, \\ 1 \; if \; Target = unvoiced \; stop \; \& \; d_{vc} = 1, \\ 0 \; otherwise. \end{cases} \tag{5}$$

## 4. Results

The detection of GAE is illustrated in Figures 5 and 6. Figures 5(a)-(c) represent the speech waveform, $r_{zb}$ and $d_{lfdvr}$ for the word [dada] uttered by normal speaker. The evidence, $d_{lfdvr} = 1$ at consonant region indicates the presence of glottal activity for the target voiced stop. This indicates the absence of GAE, i.e., GAE=0. Figures 5(d)-(f) show the speech, $r_{zb}$ and $d_{lfdvr}$ for the response of CLP speaker for the target word [dada]. Here, $d_{lfdvr} = 0$ for the initial consonant indicates the absence of glottal vibrations. This case is considered as the presence of GAE for the target voiced stop i.e., GAE=1. Figures 6(a)-(c) illustrates the analysis of GAE for word [papa] produced by the normal speaker. The presence of nasal substitution for the unvoiced stop is detected as the presence of GAE (GAE=1), which is demonstrated in Figures 6(d)-(f). The proposed algorithm is evaluated over the database described in Section 2. The syllable nuclei detection is a crucial step in the proposed GAE detection algorithm. Therefore, the estimated syllable nuclei positions are evaluated using the manually marked vowel boundaries. If the detected syllable nucleus is present within the manually marked vowel region, then it considered as correctly detected syllable nucleus. The syllable nuclei detection algorithm is evaluated for the syllables of normal and CLP speakers. Syllable nuclei detection rate is found to be 100% for the current database.

The detection of GAE is carried out at syllable level. Hence, the detected GAE i.e., $\widehat{GAE}$ for each syllable is evaluated against the ground truth derived from the visual observation of waveform using PRAAT. Table 2 shows the detection rate of GAE=0 and GAE=1 for voiced and unvoiced target stops produced by normal and CLP speakers. Stops produced by controlled normal group do not possess any GAE, hence, only detection rate of GAE=0 is reported. Whereas for CLP speakers, detection rates of cases: GAE=0 and GAE=1 are reported. The overall accuracy of proposed system is found to be $88.96\%$ and $92.33\%$ for the target voiced and for unvoiced stops, respectively.

Table 2: *Detection rates (DR) of glottal activity error present (GAE=1) and absent (GAE=0) for voiced and unvoiced target stops.*

| | Target voiced stop | Target unvoiced stop |
|---|---|---|
| Group | DR (%) | DR (%) |
| Normal (GAE=0) | 97.21 | 97.90 |
| CLP (GAE=0) | 88.37 | 95.13 |
| CLP (GAE=1) | 81.30 | 85.95 |
| Average | 88.96 | 92.33 |

## 5. Conclusion

In this work, a signal processing based algorithm is proposed for the automatic detection of GAE during the production stop consonants in speakers with CLP. The low-frequency dominant voiced consonant evidence derived from the ZFFS and BPFS is used to detect the GAEs. The detected GAEs are evaluated against the ground truth derived form PRAAT based waveform analysis. The proposed algorithm gives the information about the deviant glottal source mechanism during the production of stops. Hence, the GAE detection algorithm can be used as an assistive tool by SLPs for the better assessment of articulation errors produced by speakers with CLP.

## 6. Acknowledgement

This work is in part supported by the project grants, for the projects entitled “NASOSPEECH: Development of Diagnostic system for Severity Assessment of the Disordered Speech” funded by the Department of Biotechnology (DBT), Govt. of India and “ARTICULATE +: A system for automated assessment and rehabilitation of persons with articulation disorders” funded by the Ministry of Human Resource Development (MHRD), Govt. of India.

## 7. References


[1] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE signal processing letters*, vol. 16, no. 6, pp. 469–472, 2009.

[2] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[3] A. Harding and P. Grunwell, "Characteristics of cleft palate speech," *International Journal of Language & Communication Disorders*, vol. 31, no. 4, pp. 331–357, 1996.

[4] J. E. Trost, "Articulatory additions to the classical description of the speech of persons with cleft palate," *The Cleft palate journal*, vol. 18, no. 3, pp. 193–203, 1981.

[5] G. Henningsson, D. P. Kuehn, D. Sell, T. Sweeney, J. E. Trost-Cardamone, and T. L. Whitehill, "Universal parameters for reporting speech outcomes in individuals with cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 45, no. 1, pp. 1–17, 2008.

[6] F. Y. Al-Tamimi, A. I. Owais, O. F. Khabour, and Z. A. Khamaiseh, "Phonological processes in the speech of jordanian arabic children with cleft lip and/or palate," *Communication Disorders Quarterly*, vol. 32, no. 4, pp. 247–255, 2011.

[7] B. J. Philips and R. D. Kent, "Acousticphonetic descriptions of speech production in speakers with cleft palate and other velopharyngeal disorders," *Speech and Language*, vol. 11, pp. 113 – 168, 1984. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780126086119500085

[8] A. Maier, F. Hönig, T. Bocklet, E. Nöth, F. Stelzle, E. Nkenke, and M. Schuster, "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.

[9] L. He, J. Zhang, Q. Liu, H. Yin, and M. Lech, "Automatic evaluation of hypernasality and consonant misarticulation in cleft palate speech," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1298–1301, 2014.

[10] J. R. Orozco-Arroyave, S. Murillo-Rendón, A. M. Álvarez-Meza, J. D. Arias-Londoño, E. Delgado-Trejos, J. Vargas-Bonilla, and C. G. Castellanos-Domínguez, "Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[11] 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.

[12] A. K. Dubey, S. R. M. Prasanna, and S. Dandapat, "Zero time windowing analysis of hypernasality in speech of cleft lip and palate children," in *IEEE Twenty Second National Conference on Communication (NCC)*, 2016, pp. 1–6.

[13] L. He, J. Zhang, Q. Liu, J. Zhang, H. Yin, and M. Lech, "Automatic detection of glottal stop in cleft palate speech," *Biomedical Signal Processing and Control*, vol. 39, pp. 230–236, 2018.

[14] P. Boersma *et al.*, "PRAAT, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[15] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.