Interspeech 2018
2-6 September 2018, Hyderabad

# Processing Transition Regions of Glottal Stop substituted /s/ for Intelligibility Enhancement of Cleft Palate Speech

*Protima Nomo Sudro*[1], *Sishir Kalita*[1], *S R Mahadeva Prasanna*[1,2]

[1]Indian institute of Technology Guwahati, Guwahati, India
[2]Indian institute of Technology Dharwad, Dharwad, India

(protima,sishir,prasanna)@iitg.ernet.in

## Abstract

The speech intelligibility of cleft palate (CP) individuals is degraded primarily due to compensatory articulation errors and hypernasality. The present work proposes a method to enhance the CP speech intelligibility, where fricatives are substituted by compensatory articulation errors. Apart from the distortion present in the sustained fricative region, the fricative-vowel and vowel-fricative regions are also deviated due to co-articulation effect. Since important perceptual cues are embedded in the transition regions. Therefore, it is necessary to enhance the transition regions for more intelligibility. Motivated by the perceptual significance of the transition regions, 2D-DCT based joint spectro-temporal features are exploited for the modification. The 2D-DCT coefficients of CP speech are modified by projecting them onto the singular vectors derived from the SVD analysis of normal speech. Further, for the evaluation of speech intelligibility, objective and subjective assessment is conducted. The results show significant improvement in the speech intelligibility of the modified speech.

**Index Terms**: intelligibility enhancement, CP speech, glottal stop, transition region, joint spectro-temporal feature

## 1. Introduction

Speech intelligibility is very important for the communication in day-to-day life. However, the degradation in the intelligibility due to distorted speech patterns of the cleft palate (CP) individuals create difficulty while communicating with others. Generally, individuals with CP exhibit several speech-related disorders, such as hypernasality, compensatory articulation errors, and nasal air emission, which occurs due to the velopharyngeal dysfunction (VPD) and oro-nasal fistula. The compensatory articulation patterns have a more detrimental effect on the speech intelligibility, compared to the other speech disorders [1, 2]. The maladaptive compensatory articulations mainly affect the pressure consonants, such as fricatives, stops, and affricates, as the adequate intra-oral pressure required for these sounds are not attainable by the CP individuals [3]. Glottal stop is one such compensatory articulation, where the place of articulation of target pressure consonant is shifted to the laryngeal level, and it is associated to a greater degree of VPD. The silence like characteristics preceding the vowel and deviant formant transitions in the fricative-vowel (FV) transition region are the typical characteristic of the glottal stop substitution [4]. Among other types of compensatory articulation, glottal stop has a more severe impact on the speech intelligibility of CP speech [2]. The substitution of a fricative /s/ sound by a glottal stop is due to shift in place of articulation as well as the distorted manner of articulation [5, 6]. For this kind of phoneme specific misarticulation, clinical intervention like speech therapy may be used to improve its intelligibility [7]. Additionally, the therapy may be more effective if the enhanced version of the misarticulated target stimulus is demonstrated to the CP individuals. The improved speech can be considered as a benchmark by the clinicians while making decisions regarding the need for further surgical or prosthetic treatment. To improve the strongly habituated pattern of compensatory articulation using speech therapy, the demonstration of the enhanced version of speech may provide more impact on the proper speech development of CP individuals. The SLPs can use the enhanced speech signal in auditory discrimination test, to observe how the children with CP perceive the enhanced versions of their own recorded deviant speech [8].

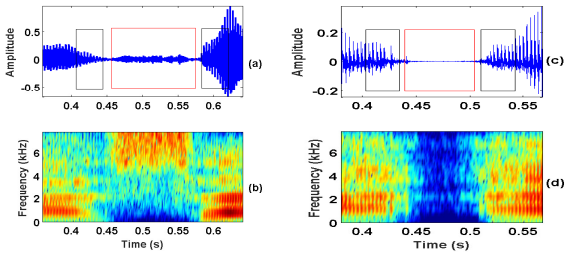


Figure 1: *Illustration of the waveform and respective spectrogram of a fricative in the intervocalic context /asa/ of a normal speaker (a & b) and CP speaker (c & d).*

Despite the possible advantages of the misarticulated CP speech enhancement, no attempts have been made in literature towards this direction. However, several works have been reported for the intelligibility enhancement of other pathological speech, which includes alaryngeal speech enhancement, dysarthric speech transformation, glossectomy speech enhancement, consonant enhancement for articulation disorder resulting from cerebral palsy, and electrolaryngeal speech enhancement [9, 10, 11, 12]. The motivations behind the intelligibility enhancement of the pathological speech, as mentioned above are to assist them with speaking aids, adapt better communication skills, etc.

In this work, the intelligibility enhancement of alveolar fricative /s/, substituted by glottal stop is carried out in fricative-vowel-fricative-vowel (FVFV) and vowel-fricative-vowel (VFV) contexts. Fig. 1 shows the characteristic differences between fricative /s/ substituted by glottal stop and normal fricative /s/ for the /asa/ context. The substitution of silence like region due to glottal stop misarticulation for the fricative /s/ can be observed from Fig. 1(c) (red rectangle). The concentration of spectral energy in the higher frequency region as observed for normal /s/ is not observed in the glottal stop substitution (Fig. 1(b & d)). Apart from the distortion in the sustained frication region, the significant deviations in the FV and



10.21437/Interspeech.2018-1646

VF transition regions are also observed in Fig. 1(d) (corresponding to black boxes in Fig. 1(c)). As a result, the abrupt transition region is observed between the fricative /s/ substituted by glottal stop and adjacent vowel. It is because there occurs no vocal tract constriction prior to the opening of vocal folds upon the release of glottal stop. The transition region between two sounds carries important perceptual cues related to the speech intelligibility [13]. Hence, the modification of the transition region is essential along with the sustained frication region for more intelligible speech.

Thus, this work proposed a method for the intelligibility enhancement of CP speech by modifying the sustained fricative region and the corresponding transition regions. The sustained frication region is transformed using insertion method, and transition regions are modified by projection method. As the transition region exhibit dynamic spectral and temporal cues, features which properly model the spectro-temporal modulations need to be considered for the intelligibility enhancement [14]. In this work, 2D-DCT based joint spectro-temporal features are exploited for modeling the transition regions. The 2D-DCT coefficients of CP speech are modified by projecting them onto the normal speech based on SVD analysis.

Rest of the paper is organized as follows. Section 2 describes the database and methods exploited for the modification of the misarticulated error i.e., insertion and 2D-DCT based joint spectro-temporal modeling of the transition region and its modification. In Section 3 subjective and objective evaluation results are discussed followed by conclusions and future directions in Section 4 .

## 2. Database and Methodologies

### 2.1. Database

The database used for the modification of glottal stop substituted /s/ correspond to the Kannada language. Speech materials used in this work are collected from All India Institute of Speech and Hearing (AIISH), Mysore, India. These materials are obtained from 22 normal speakers and 32 CP speakers. The age of CP and non-CP participants are $9 \pm 2$ years (mean $\pm$ SD) and $10 \pm 2$ years (mean $\pm$ SD), respectively. Not one of the CP participant bears any history of hearing impairment as well as any developmental difficulties. The speech material of the CP database used in this work consists of manually annotated sequence of phoneme labels with time alignments. Three repetitions of each of the word in FVFV and VFV contexts from each speaker is recorded. The signals are recorded at 48 kHz and downsampled at 16 kHz for analysis and modification. All the speech samples used in this work are auditorily judged to be glottal stops by three expert speech-language pathologists (SLPs) and also the corresponding spectrograms are visually analyzed using PRAAT software.

### 2.2. Transformation of Glottal stop in CP speech

For the analysis and modification of the fricative /s/ sound, the signal is assumed to be a combination of sustained fricative region and transition region components. The sustained region represents the turbulent fricative portion. The other component represents the transition between fricative & vowel and vowel & fricative. The overall block diagram of the modification method is shown in Fig. 2. For the transformation, each of the sustained and transition region component is modified independently. At first, the sustained region is modified using the insertion method. Next, the Mel-log time-frequency representation (Mel-log TFR) of the transition region component is modeled using 2D-DCT. The 2D-DCT coefficients of CP speech are modified by projecting them onto the singular vectors of normal speech. The singular vectors of the projection matrix are attained by SVD analysis of the centered covariance matrix of 2D-DCT coefficients of transition region components of normal speech. Finally, both the modified components are combined to obtain the target speech signal.

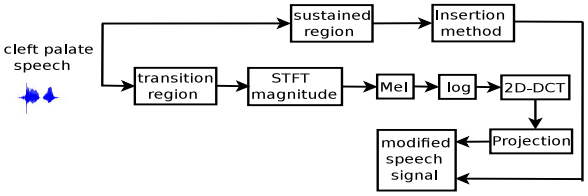


Figure 2: *Overview of the speech modification method*

#### 2.2.1. Insertion method for sustained region modification

In the insertion method, fricative /s/ sound is synthetically generated using the knowledge of normal /s/ acoustic characteristics. For the generation of /s/ sound, a white noise source with zero mean and unit variance is passed through a bandpass filter. The cut-off frequencies for the bandpass filter are 4 & 7.5 kHz respectively. The synthesized source signal of /s/ is forward filtered with average linear prediction (LP) coefficients obtained from normal speech. The resultant signal is approximated to have normal like /s/ characteristics. Then, the synthetically generated /s/ is inserted in the glottal stop substituted /s/ region. The signal after modification is shown in Fig. 3. Although, the transformed signal shown in Fig. 3 is observed to be more intelligible due to the presence of energy in the fricative portion compared to the original speech signal shown in Fig. 1 (c) & (d). However, the intelligibility of fricative /s/ does not depend only on the spectrum of steady-state portion of the sound. It also depends on the dynamic formant characteristics of adjacent FV and VF transition regions and relative frequency of fricative with that of the vowel [15]. The absence of one of the additional cues for the fricative identification are observed in the spectrogram of Fig. 3 in the form of abrupt transition region (fricative-vowel and vowel-fricative) components. Here, no formant transitions are observed in the adjacent vowel. These transitional components in speech represent the articulator gestures when they move from one position to another. The abrupt transitions are prominently observed by calculating the rate of FV/VF transition using the function $f_{FV}/f_{VF}$ as noted in Refs. [16] for each frame of the utterance. The rate of FV and VF transition for each frame $i$ for a particular FV and VF occurrence is given by,

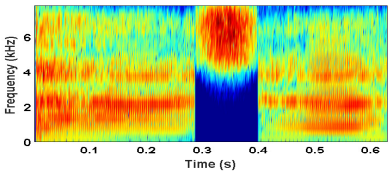


Figure 3: *Illustration of the utterance /asa/ by an individual with CP after modification using insertion method*

$$f_{FV}(i) = d(\bar{x}_f, x_i) - d(\bar{x}_v, x_i)$$
$$f_{VF}(i) = d(\bar{x}_v, x_i) - d(\bar{x}_f, x_i) \quad (1)$$

where $\bar{x}_f$ and $\bar{x}_v$ are the mean MFCC vectors for fricative and vowel in a given FV occurrence by averaging feature vectors over all frames in each phone. The variable $x_i$ denotes the feature vector of $i^{th}$ frame and d denotes the Euclidean distance between two feature vectors. An example of the FV and VF

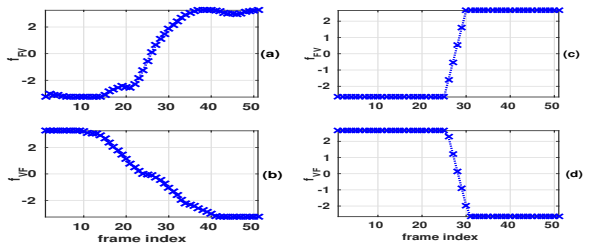


Figure 4: *Comparison of rate of FV and VF transition region components for, (a,b) normal and, (c,d) CP speech*

pair from the word /asa/ is shown in Fig. 4. It is observed that the slope of the transition region in the modified speech utterance (Fig. 4 (c) & (d)) is very low as compared to the slope from the transition region of normal speaker in Fig. 4 (a) & (b) for both FV and VF regions. The steep transition regions may be due to the abrupt opening and closing of the vocal folds lacking gradual formant transitions in the adjacent vowel. Further, the modified fricative /s/ is analyzed for fricative-to-vowel (FV) intensity ratio, FV and VF transition duration to obtain a quantitative measure. The FV intensity ratio is computed to measure the difference between original and modified speech. The FV ratio is the ratio of fricative and vowel intensity. Various researchers have demonstrated the relation between speech intelligibility and consonant-vowel (CV) intensity ratio (in this work, CV ratio is denoted as FV ratio). The FV ratio is reported to be lower for less intelligible speech [17]. This observation is in agreement with the FV intensity ratio values tabulated in Table. 1. The glottal stop substituted /s/, i.e., the unmodified/original CP speech utterance has lowest F-V intensity ratio. On the contrary, normal speech has higher F-V ratio inferring it to be more intelligible. The shorter FV and VF transition durations as compared to normal speech affirm abrupt transitions from F to V and vice versa. Hence, to further improve the intelligibility of the signal, the transition regions are required to be modified in addition to steady-state/sustained portion of the sound.

Table 1: *FV intensity ratio, FV and VF transition duration for normal, unmodified CP (UCP) and modified CP (MCP) speech utterances. F and V refer to fricative and vowel respectively*

| Measures | Normal | UCP | MCP |
|---|---|---|---|
| FV intensity ratio (dB) | -12.88 | -24 | -8.9 |
| FV transition (ms) | 0.2 | - | 0.07 |
| VF transition (ms) | 0.3 | - | 0.08 |

### 2.2.2. *Projection method for transition region modification*

To further improve the speech intelligibility, the transition region modification is performed. The transition region is first modeled using 2D-DCT [18] followed by SVD analysis and projection.

The short time Fourier transform (STFT) magnitude is computed for the dynamic transition region. Then Mel-log energies computed for each STFT magnitude spectrum is stack temporally, and the resultant matrix is termed as Mel-log TFR. Then Mel-log TFR is divided into overlapping patches of a certain size. On each of the patch, 2D-DCT is applied to capture the dynamic characteristics of the transition region [19]. For 2D-DCT modeling, the patch is represented as a matrix, $z = f(x, y)$ of N points, where, $N = PQ$. Here, $P$ denotes the spectral extent, which corresponds to the number of Mel-filter bank. While $Q$ denotes the temporal extent and it represents the number of frames which cover 40 ms. 2D-DCT of patch $f(x, y)$ is computed as

$$F(k,l) = \frac{2w(k)w(l)}{\sqrt{(PQ)}} \sum_{x=0}^{Q-1} \sum_{y=0}^{P-1} f(x,y) cos \frac{\pi l(2x+1)}{2Q} \times cos \frac{\pi k(2y+1)}{2P} \quad (2)$$

where, $w(k) = \begin{cases} \frac{1}{\sqrt{2}} & ,k = 0 \\ 1 & ,k \neq 0 \end{cases}$
and $k = 0, 1, ...P-1, l = 0, 1, ....Q-1$. The low order coefficients of 2D-DCT are retained in order to obtain a smooth approximated $\hat{f}(x, y)$ for the spectro-temporal envelope $f(x, y)$. Then, SVD analysis is performed on the 2D-DCT coefficients [20, 21]. For the given Mel-scale filter bank output (log spectrum), $x_k$ at $k^{th}$ frame, the covariance matrix $C_{xx}$ is defined as,

$$C_{xx} = \frac{1}{N} \phi(x_k) \phi(x_k)^T \quad \text{where, } \phi(x_k) = \phi(x_k) - \frac{1}{N} \sum_{k=1}^{N} \phi(x_k) \quad (3)$$

The centered covariance matrix $C_{xx}$ is decomposed into singular value decomposition (SVD) as,

$$C_{xx} = Q \Delta Q^T \quad (4)$$

where $Q \& Q^T$ are matrices of left and right singular vectors respectively. $\Delta$ represents the diagonal matrix of singular values. The 2D-DCT coefficients of the transition regions of CP speech are transformed by rotating the coefficients using right singular vectors of the projection matrix, $Q^T$. The modification/rotation is performed by multiplying the 2D-DCT coefficients of CP speech with the right singular vectors of normal speech attained using SVD analysis [22, 23].

For reconstruction of the signal, modified patches are obtained by applying inverse 2D-DCT which are then added in an overlapping manner. The time segments are merged to obtain the full sequence of STFT coefficients. Then, an inverse Fourier transform is applied to the spectrogram of each time segments from which the signal is synthesized using an overlap-add method. The modified transition regions are combined with the modified sustained fricative region of the speech to obtain an intelligible speech signal. Fig. 5 shows four different spectrogram of /asa/ segment. The silence region is modified by inserting the synthetic /s/ in Fig. 5(c) and in Fig. 5(d) combined result of modified sustained and transition region is shown. It can be observed that Fig. 5(d) is possessing normal like characteristics as shown in Fig. 5(a). The resultant files are evaluated using perceptual and objective measures. Through the combination of synthetic fricative /s/ and modified transition regions, it is approximated to have normal like intelligible speech.

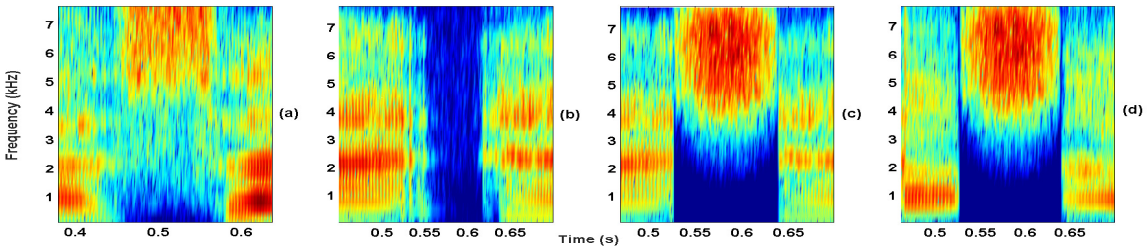


Figure 5: *Illustration of the spectrogram of /asa/ segment from the word /sasa/ for, a) normal speech, b) original CP speech (fricative /s/ is substituted by glottal stop), c) modified sustained /s/ using insertion method and d)modified sustained and transition region components of fricative /s/*

## 3. Evaluation

To compare the speech intelligibility of modified speech with the original speech, objective and subjective tests are carried out. The effectiveness is objectively analyzed using metrics, namely, Mel cepstral distortion (MCD), FV and VF transition region duration.

### 3.1. Objective evaluation

For the objective evaluation, we measure the FV and VF transition regions duration after transition region modification by projection method and combining it with the modified sustained portion. The results are tabulated in Table 2. It is observed that the FV/VF transition region duration has now increased from $0.07, 0.08$ to $0.21, 0.26$, respectively. This implies that the signal characteristics have changed from that of the sustained region modified signal. The Mel cepstral distortion (MCD) is

Table 2: *FV & VF transition duration and MCD for normal speech utterance, CP speech utterance modified by insertion method (Modified 1) and CP speech utterance modified by both insertion & projection method (Modified 2)*

| Measures | Normal | Modified 1 | Modified 2 |
|---|---|---|---|
| FV transition (ms) | 0.2 | 0.07 | 0.21 |
| VF transition (ms) | 0.3 | 0.08 | 0.26 |
| MCD | - | 0.84 | 0.82 |

computed between the normal & Modified 1 (CP speech utterance modified by insertion method) and normal & Modified 2 (CP speech utterance modified by both insertion & projection method). From the MCD values, it is observed that Modified 2 utterance has lower spectral distortion compared to the Modified 1 utterance. From Table 2, it can be noticed that the speech signal after sustained and transition region modification, possesses characteristics closer to the normal speech signal.

### 3.2. Subjective evaluation

A Listening test is conducted to compare the speech intelligibility of original and modified speech. The results of comparison test are summarized in Table 3 . The original, Modified 1 and Modified 2 CP speech utterances were presented to the listeners and were instructed to choose the utterance which is tending more like an intelligible fricative /s/ out of the three versions of each word presented to them. The listeners acquire the knowledge of speech processing. A total of ten listeners have participated in this study. It has been reported in the literature that stimuli with no formant transitions are less preferred by listeners [15]. And we observe similar kind of result in this study, as the listeners preferred the Modified 2 version compared to the original and Modified 1 versions. For reference purpose, some of the speech files are available in the link:

https://drive.google.com/drive/folders/18qOTSsoTCOdMJ2LlOvB3cVP9-ziAOxaZ.

Table 3: *Results of comparison test for misarticulated /s/ by naive listeners. GS denotes that a glottal stop substitutes the fricative /s/, Modified 1 denote the speech utterances modified by insertion method and Modified 2 denote the speech utterances modified by both insertion & projection method respectively.*

| Type | Original (%) | Modified 1 (%) | Modified 2 (%) | No Preference (%) |
|---|---|---|---|---|
| GS | 0 | 26.6 | 66.6 | 6.6 |

## 4. Conclusion and future directions

In this study, intelligibility improvement of alveolar fricative /s/ substituted by glottal stop is performed, by considering the modification of the sustained and the transition regions. The absence of frication in the sustained portion of fricative is modified using insertion method. Then, the transition region is modified using projection method based on the SVD analysis of 2D-DCT coefficients of the transition region of the normal speech utterances. The proposed method is evaluated using objective and subjective assessment. The results demonstrate that modified speech provide significant improvement in speech intelligibility. In the current work, only one type of misarticulation is addressed. Our future work is planned to explore the projection method for improving the speech intelligibility for different articulation errors in CP speech. We would also like to extend our work for improving sentence-level intelligibility.

## 5. Acknowledgement

The authors would like to thank Dr. M.Pushpavathi, AIISH Mysore, for providing CLP speech samples and providing insight about CLP speech disorder. The authors would also like to thank the research scholars of Signal processing and Signal informatics and EMST lab for their participation in subjective test. This work is in part supported by a project entitled NASOSPEECH: Development of Diagnostic system for Severity Assessment of the Disordered Speech funded by the Department of Biotechnology (DBT), Govt. of India.

# 6. References


[1] D. Sell, A. Harding, and P. Grunwell, "A screening assessment of cleft palate speech (great ormond street speech assessment)," *International Journal of Language & Communication Disorders*, vol. 29, no. 1, pp. 1–15, 1994.

[2] S. J. Peterson-Falzone, M. A. Hardin-Jones, and M. P. Karnell, *Cleft palate speech*. Mosby St. Louis, 2001.

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

[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] N. Kido, M. Kawano, F. Tanokuchi, Y. Fujiwara, I. Honjo, and H. Kojima, "Glottal stop in cleft palate speech," 1992.

[6] A. W. Kummer, M. Briggs, and L. Lee, "The relationship between the characteristics of speech and velopharyngeal gap size," *The Cleft palate-craniofacial journal*, vol. 40, no. 6, pp. 590–596, 2003.

[7] M. A. Hardin-Jones and D. L. Jones, "Speech production of preschoolers with cleft palate," *The Cleft palate-craniofacial journal*, vol. 42, no. 1, pp. 7–13, 2005.

[8] S. Strömbergsson, "Synthetic correction of deviant speech–children's perception of phonologically modified recordings of their own speech," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[9] N. Bi and Y. Qi, "Application of speech conversion to alaryngeal speech enhancement," *IEEE transactions on speech and audio processing*, vol. 5, no. 2, pp. 97–105, 1997.

[10] Y.-Y. Kong and A. Mullangi, "On the development of a frequency-lowering system that enhances place-of-articulation perception," *Speech communication*, vol. 54, no. 1, pp. 147–160, 2012.

[11] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[12] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Consonant enhancement for articulation disorders based on non-negative matrix factorization," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–4.

[13] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 769–773, 1955.

[14] S. Nittrouer and D. H. Whalen, "The perceptual effects of child–adult differences in fricative-vowel coarticulation," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1266–1276, 1989.

[15] J. M. Heinz and K. N. Stevens, "On the properties of voiceless fricative consonants," *The Journal of the Acoustical Society of America*, vol. 33, no. 5, pp. 589–596, 1961.

[16] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[17] E. Kennedy, H. Levitt, A. C. Neuman, and M. Weiss, "Consonant–vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 103, no. 2, pp. 1098–1114, 1998.

[18] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4733–4736.

[19] V. Karjigi and P. Rao, "Classification of place of articulation in unvoiced stops with spectro-temporal surface modeling," *Speech Communication*, vol. 54, no. 10, pp. 1104–1120, 2012.

[20] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[21] S. Bakamidis, M. Dendrinos, and G. Carayannis, "Svd analysis by synthesis of harmonic signals," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 472–477, 1991.

[22] T. Takiguchi and Y. Ariki, "Pca-based speech enhancement for distorted speech recognition." *Journal of multimedia*, vol. 2, no. 5, 2007.

[23] A. H. Abolhassani, S.-A. Selouani, D. O'Shaughnessy, and M.-F. Harkat, "Speech enhancement using pca and variance of the reconstruction error model identification," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.