# SPEECH PROCESSING

# CLEFT LIP SPEECH ENHANCEMENT

**DONE BY**

**ROHITH G [CB.EN.U4AIE19026]**

## ABSTRACT

Hyper-nasality and mis-articulation degrade the quality of cleft lip and palate (CLP) speech. The structural and functional defects of CLP must be corrected through surgery and speech therapy, resulting in an improved speech signal. Speech-language pathologists assess the enhanced speech quality perceptually, and the results are highly biased. Our work involves the use of spectral subtraction and Wiener methods to obtain a perceptual benchmark against which the signal after surgery/therapy can be compared. To minimize the noise associated with CLP speech, spectral peak-valley augmentation is used. The evaluation findings reveal that following augmentation, the perceived quality of the CLP speech signal improves. Furthermore, the improved signal's quality is compared to the speech signal following palatal prosthesis/surgery. The increased speech signals outperform speech after prosthesis/surgery, according to the perceptual evaluation results.

## INTRODUCTION

Speech intelligibility is important for communication whether it is in human-human interaction mode or human-machine interaction mode. Due to articulatory impairment, the intelligibility of pathological speakers is degraded and it hinders them from communicating effectively as speakers without speech pathology do. Hence, researchers study to improve the intelligibility of pathological speech from a signal processing point of view. In this work, cleft lip and palate (CLP) speech enhancement is addressed. CLP is a birth disorder that affects the speech production system. The speech disorders occur even after clinical intervention due to velopharyngeal dysfunction, oronasal fistula, and mislearning. The CLP speech distortions are categorized into hypernasality, hyponasality, articulation error, and voice disorder. Among the many speech disorders associated with CLP speech, nasalization and articulation error are the primary factors that affect speech intelligibility. Hypernasality corresponds to a resonance disorder, and the presence of nasal resonances during speech production has an excessively perceptible nasal quality. Mostly, the voiced sounds are nasalized, and the nasal consonants tend to replace the obstruents due to severe hypernasality. Misarticulations are produced either due to the structural or functional disorder or both.

Suppression of formant peaks and decrease in the spectral peak-to-valley contrast are also observed in speech degradation under noisy environments and text-to-speech synthesizers. As a solution, the formant enhancement in terms of increasing formant amplitude and sharpening of formant peaks is carried out to enhance the degraded speech quality. The extra nasal formant can be detected by group delay and LP methods. Motivated by the improvement of resonance structure achieved from clinical methods, the paper aims to develop a signal processing-based algorithm for CLP speech enhancement. In particular, removal of nasal formants, enhancement of formant peaks, and suppression of valleys of the spectrum are carried out.

## LITERATURE SURVEY

Hyper-nasality and mis-articulation impair the quality of cleft lip and palate (CLP) speech. The problems caused by CLP can be corrected through surgery. However, it will not ensure that the resulting speech signal, in terms of its intelligibility will be improved significantly. Speech-language pathologists assess the speech quality obtained after prosthesis or surgery. Nevertheless, the results and evaluation metrics are significantly skewed; it also depends on the evaluator's expertise. A signal processing-based approach speech enhancement strategy is used in this paper to obtain an enhanced signal with enhanced peaks and suppressed valleys, i.e, the spectral flatness existing in the cleft frequencies is removed by performing the aforementioned method. In the first step, CLP speech is processed to remove its original nasal formant (250 Hz), and the hyper-nasality associated with CLP speech is reduced in the second stage via spectral peak-valley enhancement. Once this is done, the LP spectra of the enhanced signal and normal version for the same sentence are compared. The perceived quality of the CLP speech signal improves. The number of peaks has increased after enhancement. This is a general trend, as observed in different frames of speech. The sample speech was obtained from the All Indian Institute of Speech and Hearing (AIISH) in Mysore, India.

The dataset used for evaluation and study is the New Mexico Cleft Palate Center- Cleft Lip and Palate (NMCPC-CLP) database. It consists of continuous speech utterances from CLP subjects and normal subjects collected at New Mexico Cleft Lip and Palate Center. In conventional speech analysis, for an 8 kHz sampled speech signal, an LP order of 8-12 is preferred to capture the F1; F2; F3, and F4 information. The LP order should not be very high since it will violate the stationarity assumption of LP analysis. For hyper-nasal speech, in order to capture P0 and P1 information, the LP spectrum of order 16 (for fs = 8 kHz) is computed. Each sentence in the dataset is around 2.5-sec duration. The database contains sentences having several fricatives. Due to the narrow constriction, high pressure is built up in oral tact, and most of the air leaks through the nose in subjects with CLP.

## DATASET

The speech has been recorded from several patients before and after the diagnosis of the subjects. The subjects had facial deformities and were unable to pronounce the given sentences intelligibly. The list of sentences has been given in the table given below. The recordings have been made with the stereo set, at a sampling rate of 44100 Hz. The sentences have been carefully selected, which consist of fricatives, nasals, and stops.

The database split is: A total of 73 speakers: 41 CLP subjects, and 32 normal subjects. There

are a total of 1813 sentences, 677 normal and the rest hypernasal. There are three different categories among the hypernasal subjects, mild, moderate, and severe.

Among them there are nine subjects who have been recorded before and after their diagnosis, having a total of around 150 sentences, which would help assess the developments post-diagnosis in the same subject. There are 76 different utterances which are shown in the table. They are well balanced and commonly used sentences with all different sound units. These can be further pruned for analyzing any particular sound unit. Such sentences are indeed effective to assess the subjects, as they involve good movement of articulators. All the utterances have common words that are used so that the speaker is familiar with the utterance to produce it in a natural way.

| S.No | subject | AVG score | Degree of nasalization | gender | age | dysarthria description | no:files | duration |
|---|---|---|---|---|---|---|---|---|
| 1 | AC | 2.52 | severe | F | 13 | UCL and P | 14 | 17.14 |
| 2 | AH | 1.24 | mild | M | 8 | | 50 | 69.35 |
| 3 | BQ | 2.76 | severe | F | 4 | moderate to severe VPI | 10 | 16.68 |
| 4 | CB | 1.5 | moderate | F | 7 | BCL and p | 19 | 35.45 |
| 5 | CL | 0.98 | mild | F | 13 | UCL and P | 35 | 58.86 |
| 6 | CM | 0.5 | mild | M | 12 | CL,residual VPI | 12 | 17.32 |
| 7 | CN | 1.2 | mild | M | 7 | sub mucous CP | 38 | 73.35 |
| 8 | CW | 2.96 | severe | M | 8 | | 29 | 53.97 |
| 9 | DS | 1.3 | mild | M | 9 | | 30 | 50.55 |
| 10 | DV | 2.52 | severe | M | 15 | BCL and P moderate VPI | 19 | 31.46 |
| 11 | EC | 2.46 | severe | M | 8 | left ULC and P severe VPI | 25 | 31.08 |
| 12 | EC1 | 1.12 | mild | M | 7 | | 33 | 54.44 |
| 13 | EM | 1.3 | mild | M | 18 | BCL and P | 28 | 38.04 |
| 14 | ES | 1.5 | moderate | F | 23 | right UCl and P | 20 | 30.75 |
| 15 | ES1 | 0.92 | mild | F | 10 | complex VPI | 40 | 75.28 |
| 16 | GC | 2.8 | severe | F | 16 | | 38 | 61.03 |
| 17 | GE | 2.64 | severe | M | 13 | BCL and P | 24 | 37.90 |
| 18 | GR | 2.02 | moderate | F | 10 | left UCL and P | 30 | 48.42 |
| 19 | HSB | 3 | severe | F | 5 | High grade submucous CP | 38 | 71.56 |
| 20 | IC | 0.6 | mild | M | 19 | BCL and P | 36 | 49.14 |
| 21 | IL | 2.42 | moderate | M | 9 | sub mucous CP severe VPI | 39 | 55.76 |
| 22 | JA10 | 2.94 | severe | M | 9 | left UCL and P | 17 | 18.71 |
| 23 | JA2011 | 0.8 | mild | M | 7 | | 38 | 69.24 |
| 24 | JB | 2.4 | moderate | F | 6 | BCL and palate severe VPI | 28 | 53.79 |
| 25 | JB1 | 1.74 | moderate | F | 10 | BCL and palate severe VPI | 20 | 34.26 |
| 26 | JD | 3 | severe | M | 5 | BCL and palate severe VPI | 71 | 136.54 |
| 27 | JG8 | 2.74 | severe | F | 8 | CP | 37 | 71.49 |
| 28 | JL | 3 | severe | F | 5 | sub mucous CP | 15 | 22.59 |
| 29 | JM | 2.74 | severe | M | 7 | CP and severe VPI | 29 | 54.21 |
| 30 | LS | 2.18 | moderate | F | 33 | sub mucous CP | 36 | 51.27 |
| 31 | MO | 1.74 | moderate | F | 17 | Velocardiofacial syndrome | 25 | 39.89 |
| 32 | MV | 1.56 | moderate | F | 11 | VPI | 23 | 35.75 |
| 33 | PC | 1.72 | moderate | M | 11 | Wildervank syndome ,CP | 22 | 41.09 |
| 34 | RD | 1.9 | moderate | F | 15 | | 12 | 13.29 |
| 35 | RM | 1.64 | moderate | F | 8 | BCL and ,m2s VPI | 12 | 14.78 |
| 36 | RT | 3 | severe | M | 5 | Right UCL and P | 9 | 10.39 |
| 37 | SV | | | M | 11 | BCL and P | 35 | 48.02 |
| 38 | SW | 2.28 | moderate | M | 7 | Left UCL and P | 35 | 57.15 |
| 39 | SW1 | 3 | severe | F | 6 | sub mucous CL | 10 | 22.77 |
| 40 | TO | 2.54 | severe | M | 5 | BCL and P | 21 | 29.27 |
| 41 | UG | 0.84 | mild | M | 10 | | 37 | 52.96 |
| 42 | ZD | 2.42 | moderate | M | 11 | VPI | 32 | 70.42 |

## TABLE FOR ALL THE UTTERANCES FROM THE DATABASE

| S.no | **Sentences** | S.no | **Sentences** |
|---|---|---|---|
| **1** | put the baby in the buggy | **39** | sissy sees the stars |
| **2** | bobby and bill play ball | **40** | she went shopping |
| **3** | bob is a baby boy | **41** | shine the shoes |
| **4** | buy baby a bib | **42** | wash the shoes |
| **5** | papa plays base ball | **43** | see the busy bees |
| **6** | billy bob will play ball | **44** | this is tuesday |
| **7** | baby bottle baby bottle | **45** | the zebra lives at zoo |
| **8** | papa will play ball | **46** | i like cheese pizza |
| **9** | pop the big bubble | **47** | sixty six, sixty six, sixty six |
| **10** | the puppy will pull a rope | **48** | she goes to the show |
| **11** | the puppy plays with the rope | **49** | jhonny told a joke |
| **12** | take teddy to town | **50** | jack told a joke |
| **13** | do it for daddy | **51** | jimme and charlie chew gum |
| **14** | today is a good day | **52** | the children watch soccer match |
| **15** | give kate a cookie | **53** | chase the chickens |
| **16** | get kate a cookie | **54** | choo choo train |
| **17** | go get in a wagon | **55** | will jeff chew |
| **18** | go get in a car | **56** | chocolate chip cookie |
| **19** | big red truck | **57** | check the child |
| **20** | i like cake | **58** | choo choo |
| **21** | i like ice cream | **59** | hanna hurt her hand |
| **22** | i like gold | **60** | where is the way home |
| **23** | i have five fingers | **61** | mama makes muffins |
| **24** | i have a fire fly | **62** | no one is comming |
| **25** | feed the fish | **63** | mary knew no one |
| **26** | very funny | **64** | i am going away |
| **27** | i can laugh | **65** | you ran a long mile |
| **28** | vikky drives a van | **66** | i can sing a song |
| **29** | vikky drives a bus | **67** | roll the carpet |
| **30** | something smells funny | **68** | i love caramel |
| **31** | fifty fifty | **69** | thats really cool |
| **32** | i have a firefly | **70** | theach me to sing |
| **33** | very fresh fruit | **71** | i do we do you do |
| **34** | i like to laugh | **72** | aron lives in an igloo |
| **35** | the feather fell off the leaf | **73** | i stay on a train |
| **36** | i see sun | **74** | talk to teddy |
| **37** | i see the sky | **75** | dig to daddy |
| **38** | i see the stars | **76** | today will be a good day |

# METHODOLOGY

## Wiener filter

The Wiener filter is a signal processing filter that uses linear time-invariant (LTI) filtering of an observed noisy process to obtain an estimate of a desired or target random process, assuming known stationary signal and noise spectra and additive noise. Between the estimated random process and the intended process, the Wiener filter minimizes the mean square error.
The Wiener filter's purpose is to compute a statistical estimate of an unknown signal by taking a similar signal as an input and filtering it to create the estimate as an output. The known signal, for example, could be an unknown signal of interest that has been tainted by additive noise. The Wiener filter can be used to remove noise from a distorted signal and estimate the underlying signal of interest. The Wiener filter is based on a statistical method, and the minimum means square error (MMSE).

Before passing the audio signal through the Wiener filter we do the linear predictive analysis followed by the same analysis for the output obtained from the Wiener filter. For frame size of 30 ms and LP order equal to 11 we see that the output plot for both cases looks similar, i.e the number of peaks detected in both cases were equal and there is no enhancement.We go with the second methodology Spectral Subtraction.

## Spectral Subtraction

The spectral subtraction approach of noise reduction is simple and effective. An average signal spectrum and average noise spectrum are estimated in parts of the recording and subtracted from each other in this method, resulting in an increased average signal-to-noise ratio (SNR). It is assumed that the signal is corrupted by wide-band, stationary, additive noise, that the noise estimate remains constant during the analysis and restoration, and that the phase of the original and restored signals are identical.

The noisy signal y(m) is a sum of the desired signal x(m) and the noise n(m):

$$y(m) = x(m) + n(m)$$

In the frequency domain, this may be denoted as:

$$Y(j\omega) = X(j\omega) + N(j\omega) \Rightarrow X(j\omega) = Y(j\omega) - N(j\omega)$$

where $Y(j\omega)$, $X(j\omega)$, and $N(j\omega)$ are the Fourier transforms of y(m), x(m), n(m), respectively.

Because the noise's statistic characteristics are unknown, the noise and speech signal is substituted with estimates:

The time-averaged noise spectrum derived from areas of the recording where the only noise is present is used to calculate the expected noise spectrum. The following formula is used to

determine the noise estimate. The i'th of K noise frames' amplitude spectrum A first-order low-pass filter can be used to determine the noise estimate in the kth frame: The smoothed noise estimate in the i-th frame, where n is the filtering coefficient (0.5 n 0.9, but other writers prefer 0.8 n 0.95). To calculate the noise level, examine the portion of the recording that contains just noise and comes before the voice signal (the length of the analyzed fragment should be at least 300 ms). An additional speech detector must be employed to accomplish this. This inaccuracy decreases the signal quality and introduces residual noise, sometimes known as musical noise. As a result, the more noise sections are included in the study, the more accurate the noise estimate becomes. In the frequency domain, the signal-to-noise ratio can be expressed as SNR a priori (for a clean signal) or SNR a posteriori (for a noisy signal) (for a noisy signal). The SNR in the kth frame is calculated as follows:

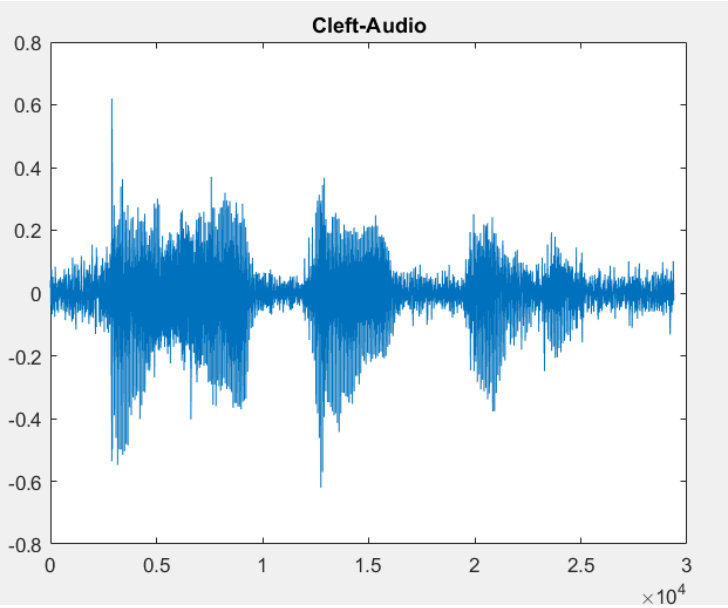
Because the clean signal is unknown throughout the restoration process, the SNR must be computed a priori. The optimal SNR in the k-th frame can be defined as follows using the Gaussian model:

The variance of the previous frame's noise spectrum is an estimate of the restored signal and $\eta$ is constant ($0.9 < \eta < 0.98$). The variance is typically substituted by a noise spectral power estimate:

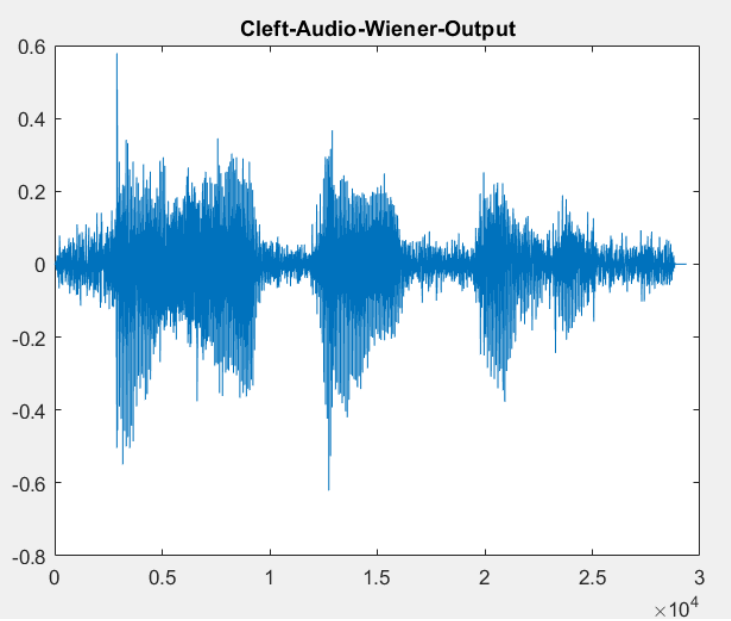
## EXPERIMENTAL CODE 1

### Wiener filter

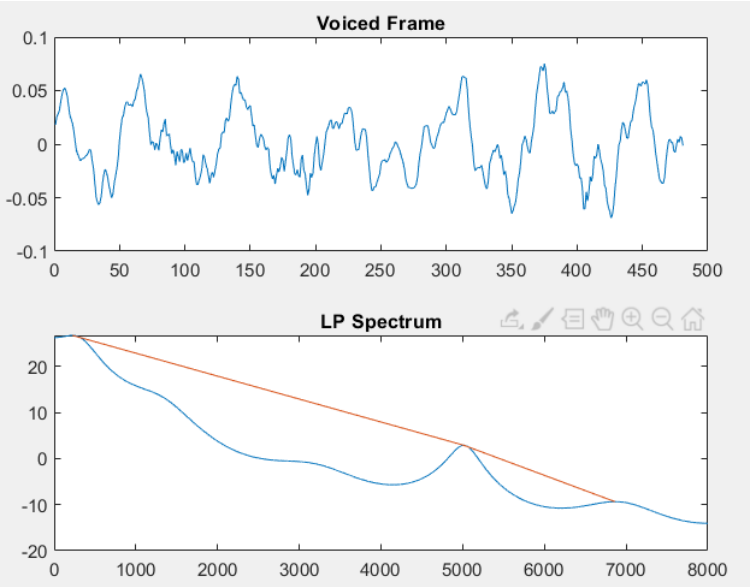
Audio file used - AC_HN_023.wav
The original cleft audio signal plot.



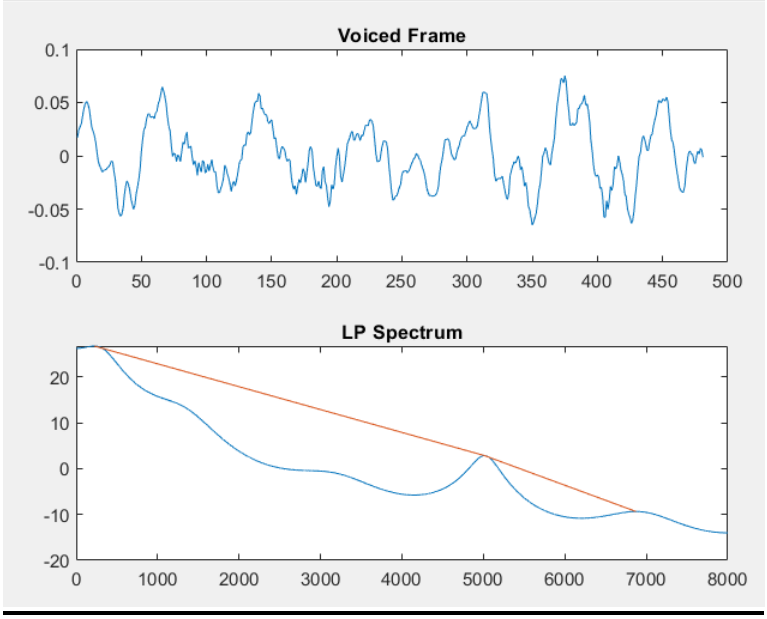

The Wiener filtered the audio signal.



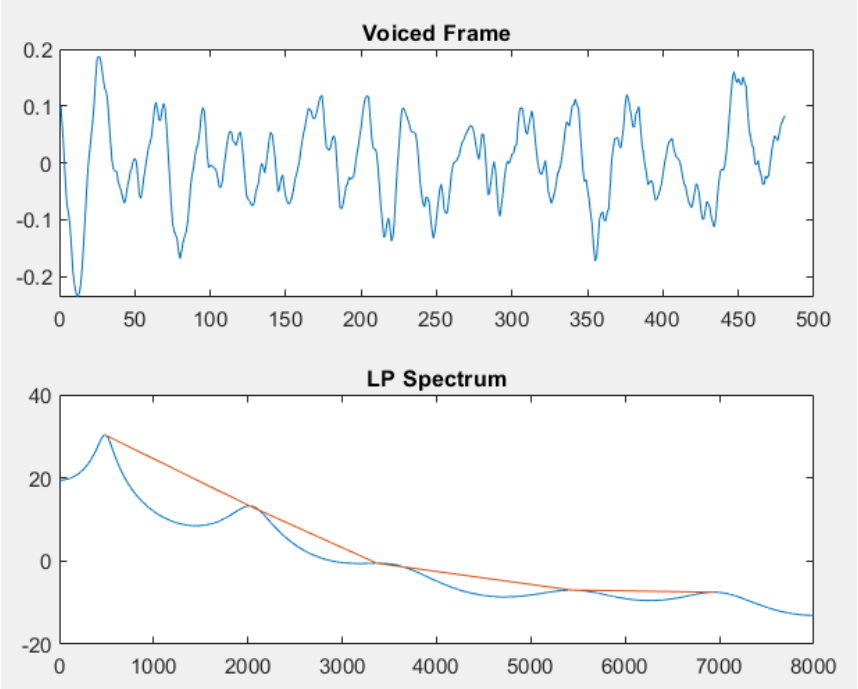

LP Spectrum for the original cleft audio file.




LP Spectrum for the Wiener filtered cleft audio file.




LP Spectrum for the Normal person's audio for the same sentence and same speaker tag.



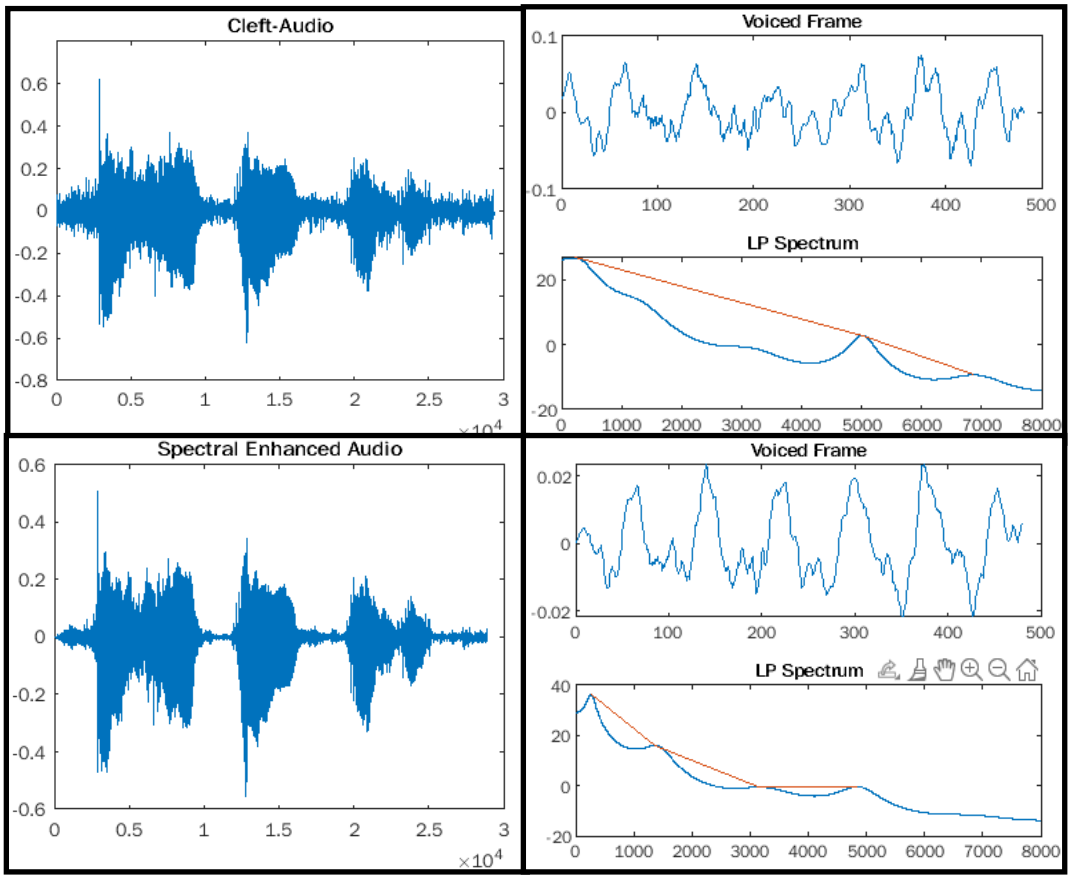
## RESULTS AND DISCUSSION

Enhancement comparison: For each of the 5 following files, the first two spectra plots give the cleft and the enhanced versions. Following the two, the spectra of the normal version of the sentence are plotted. The formant frequencies for each of the three spectra have also been given in each case.

('T_freq' denotes the starting sample of the voiced frame chosen for demonstration)

1) AC_HN_023.wav

T_freq = 10000




**Upper:** Cleft signal, Cleft voiced frame, LP spectra of the voiced frame. **Lower:** Spectrally enhanced cleft signal, Enhanced voiced frame, LP spectra of voiced frame

| Formant | In cleft signal | In enhanced | In normal speech |
|---|---|---|---|

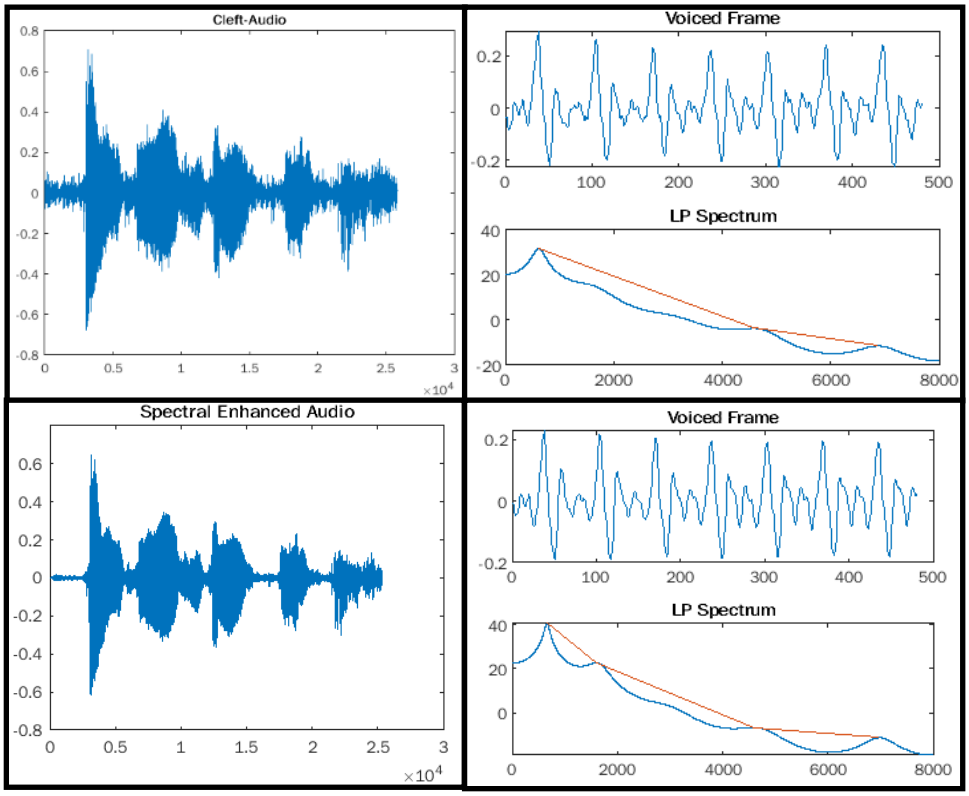
| frequencies | (in Hz) | cleft signal (in Hz) | (in Hz) |
|---|---|---|---|
| F1 | 306.5 | 269.4 | 487.9 |
| F2 | 1337.0 | 1421.7 | 2039.0 |
| F3 | 3197.5 | 3271.0 | 3575.3 |
| F4 | 5021.8 | 4879.3 | 5460.8 |
| F5 | 6932.1 | 6837.8 | 6994.5 |

**AC_NH_023.wav:** Normal version of the above sentence

Cleft: AC_HN_023.wav
Enhanced: AC_ENH_HN_023.wav
Cepstrum Distance Objective Speech Quality Measure: 3.0120
Overall SNR: 11.7526

2) AH_HN_002.wav

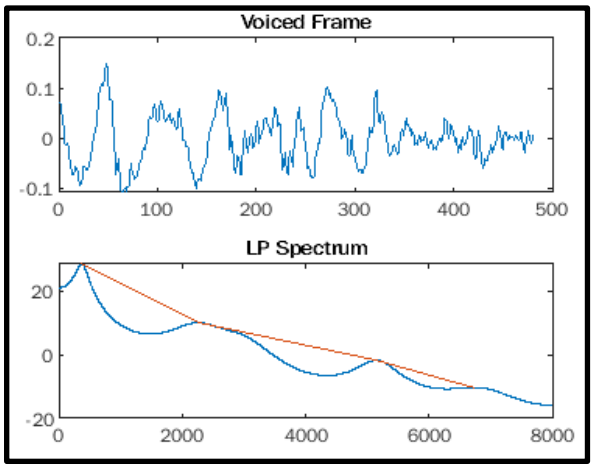
T_freq = 4600




**Upper:** Cleft signal, Cleft voiced frame, LP spectra of voiced frame. **Lower:** Spectrally enhanced cleft signal, Enhanced voiced frame, LP spectra of voiced frame

| Formant frequencies | In cleft signal (in Hz) | In enhanced cleft signal (in Hz) | In normal speech (in Hz) |
|---|---|---|---|
| F1 | 621.2 | 661.9 | 380.3 |
| F2 | 1650.8 | 1650.6 | 2229.8 |
| F3 | 3146.2 | 3006.3 | 3049.4 |
| F4 | 4723.4 | 4740.1 | 5170.1 |
| F5 | 6904.5 | 6991.2 | 6874.3 |

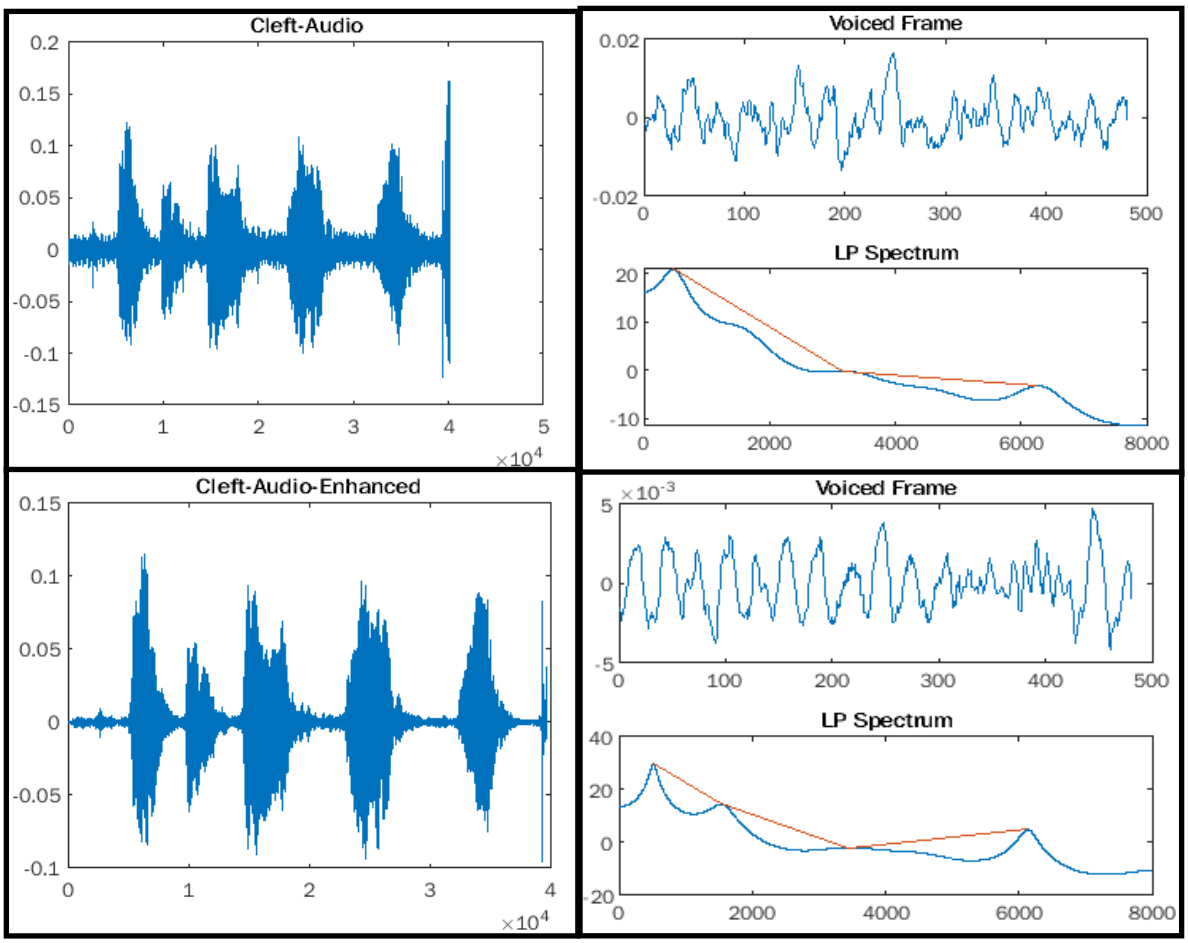
**AH_NH_002.wav:** Normal version of the above sentence




Cleft: AH_HN_002.wav
Enhanced: AH_ENH_HN_002
Cepstrum Distance Objective Speech Quality Measure: 3.4117
Overall SNR: 13.0237

## 3) BQ_HN_005.wav

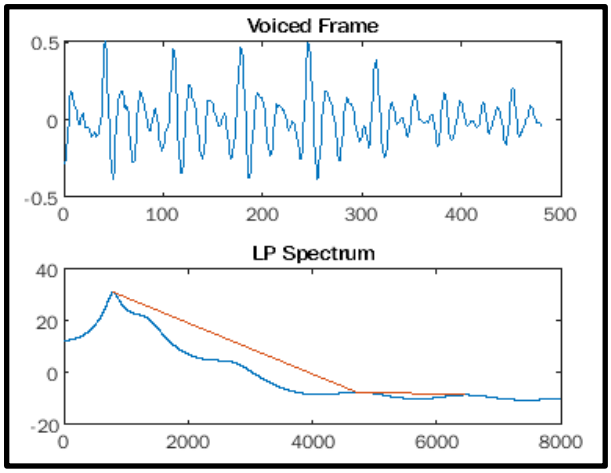
T_freq = 4600




**Upper:** Cleft signal, Cleft voiced frame, LP spectra of the voiced frame. **Lower:** Spectrally enhanced cleft signal, Enhanced voiced frame, LP spectra of voiced frame

| Formant frequencies | In cleft signal (in Hz) | In enhanced cleft signal (in Hz) | In normal speech (in Hz) |
|---|---|---|---|
| F1 | 483.1 | 517.1 | 791.0 |
| F2 | 1602.4 | 1560.5 | 1338.6 |
| F3 | 3317.2 | 3431.9 | 2745.8 |
| F4 | 4643.0 | 4537.4 | 4796.6 |
| F5 | 6314.3 | 6137.4 | 6483.9 |

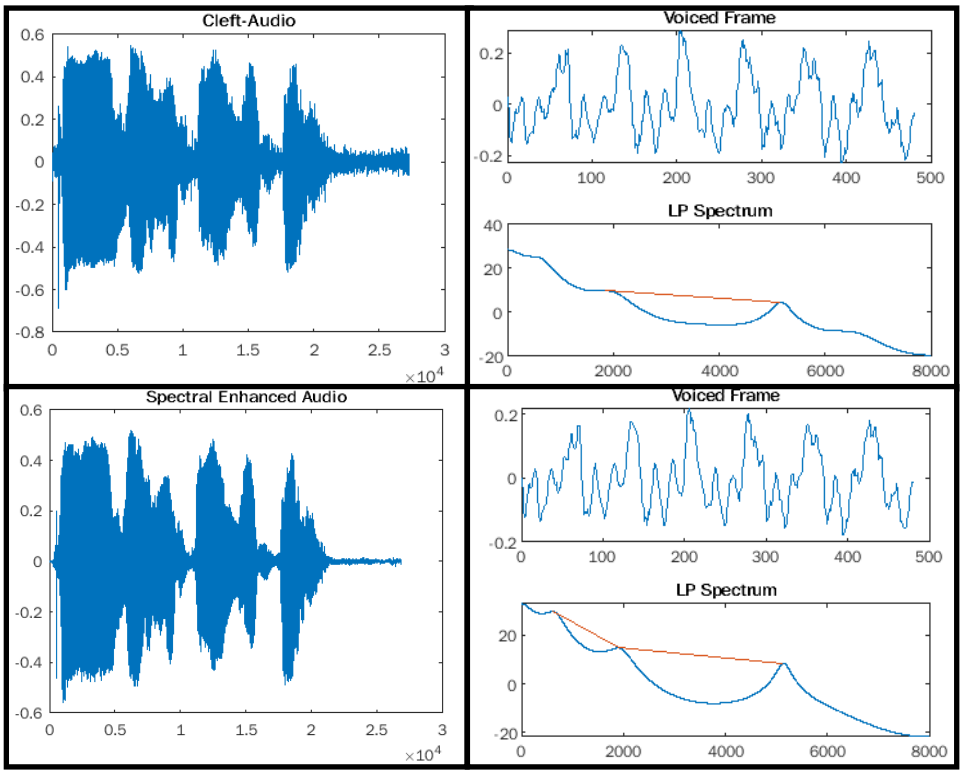
**BQ_NH_005.wav:** Normal version of the above sentence




Cleft: BQ_HN_005.wav
Enhanced: BQ_ENH_HN_005.wav
Cepstrum Distance Objective Speech Quality Measure: 2.6890
Overall SNR: 13.4415

4) CB_HN_003.wav

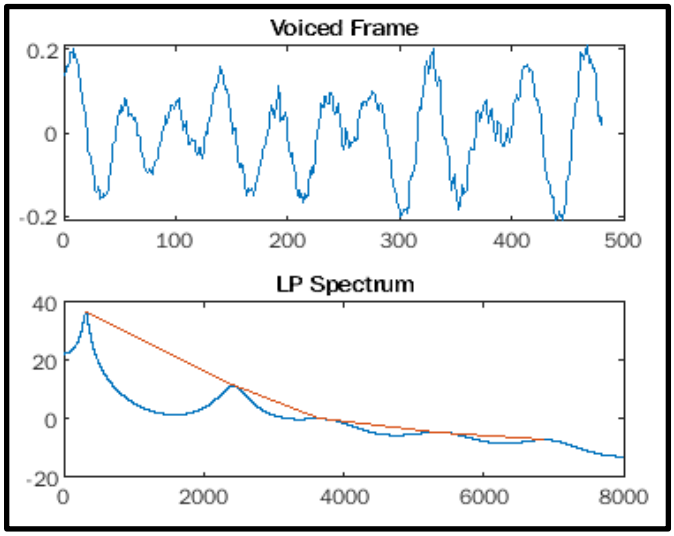
T_freq = 19500




**Upper:** Cleft signal, Cleft voiced frame, LP spectra of voiced frame. **Lower:** Spectrally enhanced cleft signal, Enhanced voiced frame, LP spectra of voiced frame

| Formant frequencies | In cleft signal (in Hz) | In enhanced cleft signal (in Hz) | In normal speech (in Hz) |
|---|---|---|---|
| F1 | 643.9 | 665.0 | 317.2 |
| F2 | 2011.7 | 1957.2 | 2422.4 |
| F3 | 3812.0 | 4199.6 | 3790.2 |
| F4 | 5177.5 | 5145.7 | 5411.0 |
| F5 | 6594.0 | 6198.8 | 6954.7 |

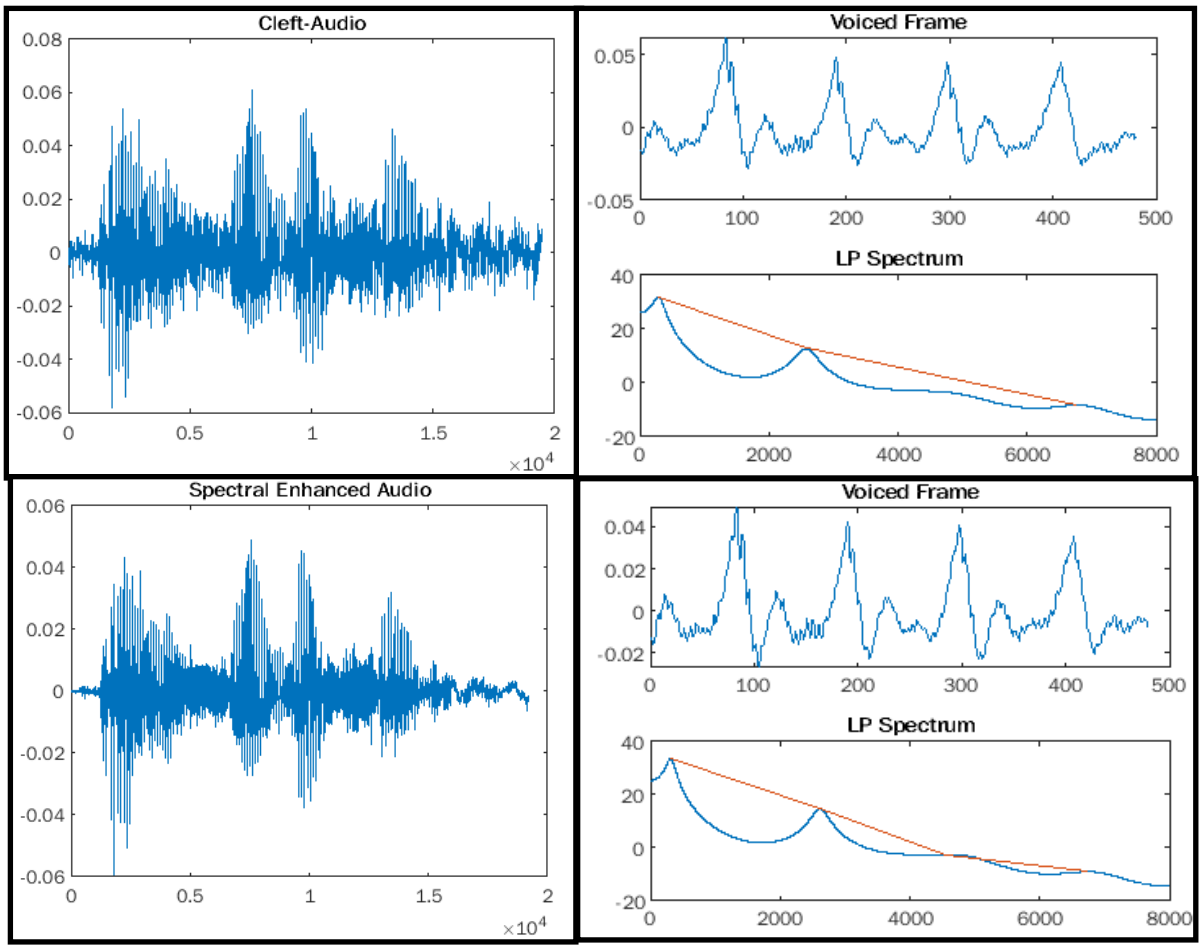
**CB_NH_003.wav:** Normal version of the above sentence




Cleft: CB_HN_003.wav
Enhanced: CB_ENH_HN_003.wav
Cepstrum Distance Objective Speech Quality Measure: 2.7499
Overall SNR: 15.3560

5) CM_HN_051.wav

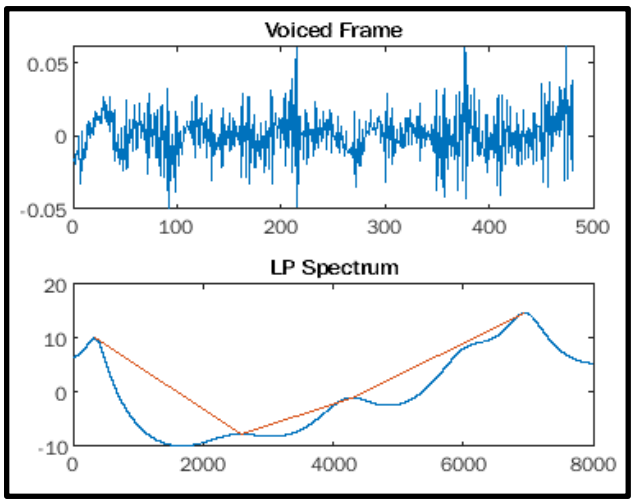
**T_freq** = 7500




**Upper:** Cleft signal, Cleft voiced frame, LP spectra of voiced frame. **Lower:** Spectrally enhanced cleft signal, Enhanced voiced frame, LP spectra of voiced frame

| Formant frequencies | In cleft signal (in Hz) | In enhanced cleft signal (in Hz) | In normal speech (in Hz) |
|---|---|---|---|
| F1 | 296.8 | 315.9 | 356.4 |
| F2 | 2580.6 | 2616.1 | 2474.0 |
| F3 | 4065.0 | 4046.9 | 4207.6 |
| F4 | 4968.0 | 4888.2 | 6003.8 |
| F5 | 6877.9 | 6858.7 | 6964.0 |

**CM_NH_051.wav:** Normal version of the above sentence




```
Cleft: CM_HN_051.wav
Enhanced: CM_ENH_HN_051.wav
Cepstrum Distance Objective Speech Quality Measure: 2.2534
Overall SNR: 8.1906
```

## CONCLUSION

Compared to the normal version of a sentence occurrence, the cleft version has less number of peaks in the LP spectra plot. Also, the difference in height between a peak and valley is less in the case of a cleft lip.

Performing spectral enhancement on the cleft signal enhanced the existing peaks and made the valleys deeper, i.e, it removed the spectral flatness in the LP spectra of the cleft plot. In some cases, new peaks also occurred on performing enhancement.

As a measure of enhancement, SNR values and cepstrum distance were also calculated.

## REFERENCES

[1] Nman DS, Thomas P, Hodgkinson PD, Reid CA, “Oro-nasal fistula development and velopharyngeal insufficiency following primary cleft palate surgery - an audit of 148 children born between 1985 and 1997,” Br. J. of Plast. Surg., vol. 58, pp. 1051–1058, 2005.

[2] Koog T, “The pharyngeal flap operation in cleft palate,” Br. J. of Plast. Surg., vol. 18, pp. 265–283,1965.

[3] Murthy J, Sendhilnathan S, Hussain S A, et. al., “Speech Outcome Following Late

Primary Palate Repair," Cleft PalateCraniofacial Journal, vol.47, no.2, pp. 156-161, 2010.

[4] Joao Henrique Nogueira Pinto, Giseleda Silva Dalben, Maria Ines Pegoraro-Krook, "Speech Intelligibility of Patients With Cleft Lip and Palate After Placement of Speech Prosthesis," The Cleft Palate-Craniofacial Journal, vol. 44, no. 6, pp. 635–641, 2007.

[5] Debbie Sell, "Issues in perceptual speech analysis in cleft palate and related disorders: a review," Int. J. Lang. Comm. Dis., vol. 40, no. 2, pp. 103–112, 2005.

[6] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Individuality preserving voice conversion for articulation disorders based on nonnegative matrix factorization," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 8037–8040.

[7] K. Xiao, S. Wang, M. Wan, and L. Wu, "Reconstruction of mandarin electrolaryngeal fricatives with a hybrid noise source," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 27, no. 2, pp. 383–391, 2019.

[8] 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.

[9] A. W. Kummer, Cleft Palate & Craniofacial Anomalies: Effects on Speech and Resonance. Nelson Education, 2013.

[10] S. J. Peterson-Falzone, M. A. Hardin-Jones, and M. P. Karnell, Cleft Palate Speech. Mosby St. Louis, 2001.