INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Nasal Air Emission in Sibilant Fricatives of Cleft Lip and Palate Speech

Sishir Kalita[1], *Protima Nomo Sudro*[1], *S. R. M. Prasanna*[1,2], *S. Dandapat*[1]

[1]Indian institute of Technology Guwahati, Guwahati, India
[2]Indian institute of Technology Dharwad, Dharwad, India
(sishir, protima, prasanna, samaren)@iitg.ernet.in

Abstract

Cleft lip and palate (CLP) is a congenital disorder of the orofacial region. Nasal air emission (NAE) in CLP speech occurs due to the presence of velopharyngeal dysfunction (VPD), and it mostly occurs in the production of fricative sounds. The objective of present work is to study the acoustic characteristics of voiceless sibilant fricatives in Kannada distorted by NAE and develop an SVM-based classification to distinguish normal fricatives from the NAE distorted fricatives. Static spectral measures, such as spectral moments are used to analyze the deviant spectral distribution of NAE distorted fricatives. As the aerodynamic parameters are deviated due to VPD, the temporal variation of spectral characteristics might also get deviated in NAE distorted fricatives. This variation is studied using the peak equivalent rectangular bandwidth (ERB_N)-number, a psychoacoustic measure to analyze the temporal variation in the spectral properties of fricatives. The analysis of NAE distorted fricatives shows that the maximum spectral density is concentrated in the lower frequency range with steep positive skewness and more variations in the trajectories of peak ERB_N-number as compared to the normal fricatives. The proposed SVM-based classification achieves good detection rates in discriminating NAE distorted fricatives from normal fricatives.

Index Terms: Cleft lip and palate, nasal air emission, spectral moments, peak ERB_N number, support vector machine.

1. Introduction

Cleft lip and palate (CLP) is one of the most common craniofacial congenital disorder. Inadequate velopharyngeal closure (velopharyngeal dysfunction (VPD)) or oro-nasal fistula (ONF) results in the oro-nasal coupling during the production of oral sounds [1, 2]. It leads to the development of several speech disorders in individuals with CLP, and one of the most common disorders is nasal air emission (NAE), which make their speech unpleasant [3, 4]. The NAE is associated with the production of obstruent sounds, and it mostly occurs with the unvoiced fricatives. Generally, during the production of fricatives, intra-oral pressure (IOP) builds up behind the narrow constriction in the vocal tract that drives the airflow through it to produce frication noise [5, 6]. However, in case of individuals with CLP, sometimes the airflow is leaked through the nasal cavity during the production of fricatives either due to the VPD or ONF. The airflow that escaped from the vocal tract system creates another turbulence noise source in the nasal cavity, which is exhaled forcibly [7]. The forceful exhalation of air from the nasal cavity while producing fricative sounds is perceived as an additional noise source, and it significantly distorts the speech intelligibility [1, 8]. This speech distortion is known as nasal air emission, and it primarily occurs in the production of pressure-sensitive phonemes. Sometimes, it may be a learned articulatory pattern to compensate for the decreased IOP.

Currently, speech-language pathologists (SLPs) evaluate the presence of NAE in fricatives by perceptual evaluation with or without the visual inspection by instruments [9, 10]. During speech therapy or any other speech intervention, SLPs examine whether NAE is perceived while accompanying or masking the consonant to evaluate the behavior of velopharyngeal port. However, the clinical reliability of perceptual evaluation depends on the expertize, which may result in a biased decision. Also, there exists a lot of variation in the NAE perception, which makes the assessment of NAE very difficult. A quantitative measure of NAE using the nasal accelerometry has been proposed in [8]. However, the instrumental assessment requires more involvement from the patients' side. David J. Zazac et. al. in [11] showed the aerodynamic and acoustic characteristics of nasal air emission for the case of fricative /s/. They have separately recorded oral and nasal signals and found that most of the spectral energies concentrate in the region of 2.5-7 kHz. However clinically, no acoustic measure is available to evaluate the NAE. In the case of CLP speech with NAE distortion, additional turbulence noise also exists, and aerodynamic parameters may deviate. It may affect the coordination among the articulators and distort the temporal variation of spectral properties of fricatives; however, these deviations are largely unknown. Moreover, no works have been reported in the literature to classify the fricatives with NAE from the normal based on speech technology-based methods. An approach based on the acoustic analysis of speech to evaluate the presence of NAE may assist SLPs with their therapeutic and other interventions.

In this work, the unvoiced sibilant fricatives /s/ and /ʃ/ distorted by NAE are considered for the analysis. A support vector machine (SVM)-based classification is proposed to distinguish the fricatives with NAE from the normal. Initially, acoustic analysis is performed to observe the spectral distribution of both the normal and NAE distorted fricatives. Static and dynamic spectral cues are investigated to characterize the fricatives with NAE. Static cues referred to the spectral characteristics for one particular speech frame, while dynamic cues are studied to analyze how the spectral properties are varying over time [12]. Four spectral moments, namely, mean (M1), standard deviation (M2), skewness (M3), and kurtosis (M4) are explored to study the static characteristics. Spectral moments are widely used to analyze the acoustic characteristic of fricatives [13, 14]. Additionally, to study the temporal variation of spectral characteristics, a method based on the analysis of ERB_N-number which denotes the dominant psychoacoustic frequency is studied [12]. In this case, gammatone filter-bank is used to compute the auditory excitation from the discrete Fourier transform (DFT) spectrum, and frequency value in ERB scale which corresponds to the maximum of auditory excitation is termed as the ERB_N-number. It is expected that the temporal variation of the ERB_N-number will be more in case of fricative with NAE. Later, from the auditory excitation, a feature is computed using

http://dx.doi.org/10.21437/Interspeech.2019-2345

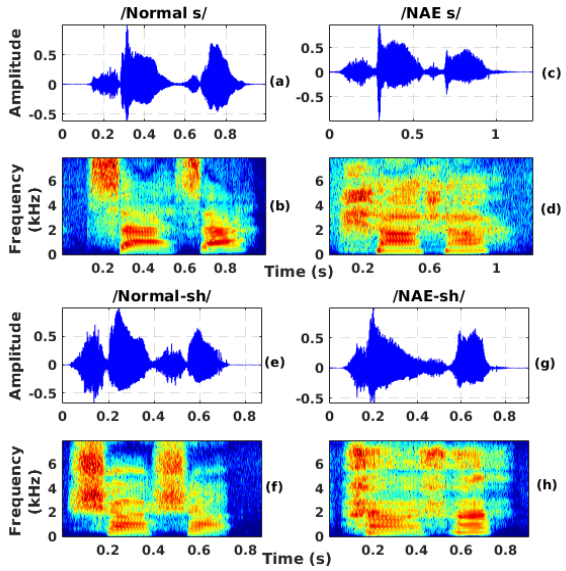

Figure 1: *Illustration of spectral deviations of normal and NAE distorted (a-d) /s/ and (e-h) /ʃ/ respectively.*

the discrete cosine transform (DCT), and this feature along with the spectral moments are used to build the SVM-based classifier to classify the fricatives with NAE from the normal fricatives.

The rest of the paper is organized as follows. Section 1 provides a description of the speech data used in this current work. The observed acoustic characteristics of the fricative with NAE are discussed in Section 2. Section 3 explained about the acoustic features extracted for the study. The detailed results and discussion of the experiments are included in Section 4. Finally, Section 5 summaries the whole work and discusses the possible directions for future work.

2. Speech data

The CLP speech utterances used in this work were obtained from All India Institute of Speech and Hearing (AIISH), Mysore, India. Speech samples of 20 CLP individuals with NAE are considered in this study. All the participants had repaired cleft of lip and palate and are native speakers of Kannada of age group 7-12 years. All the CLP individuals do not have any history of hearing impairment as well as other congenital syndrome and developmental difficulties. The language abilities of all the individuals with CLP was adequate. A group of 20 children with normal speech and language characteristics who were age and gender-matched served as controls for the study.

Generally, SLPs assess the NAE using word-level stimuli rich in pressure consonants. In this study, the fricative /s/ and /ʃ/ are collected in CVCV contexts. All the participants were seated comfortably in a soundproof room, and the data was recorded using a directional microphone kept at a distance of 15 cm from each child with a sampling frequency of 44 kHz. All the recorded speech samples are manually transcribed, and only the fricative portion is taken for the analysis.

Three experienced SLPs from the AIISH, Mysore assessed the whole database. The inter-rater reliability of the listeners are computed using the Fleiss' Kappa, and a value of 0.65 is obtained.

3. Spectral analysis of fricative with NAE

From Figure 1(c-d), it can be observed that NAE distorted /s/ have maximum spectral energy concentration ranging between 2-5 kHz compared to normal /s/ spectral energy which is concentrated above 4 kHz (Figure 1(a-b)). The normal fricative sound acquisition requires adequate IOP to create turbulence in the flow of air through the narrow constriction formed in the oral cavity. The spectral characteristics of fricative /s/ and /ʃ/ are determined by the pole/zeros created due to the resonant cavity [15]. However, in case of NAE distorted /s/, additional turbulence created in the nasal cavity lowers the required IOP for /s/ production, and its effect is evident from the spectral distortion observed in the range of 2-5 kHz. The additional turbulence generated in the nasal cavity is attributed to the coupling of oro-nasal cavity caused by the structural and functional disorder. In case of NAE distorted /ʃ/ shown in Figure 1(g)-(h) the maximum spectral energy is ranging from 1.5 kHz to 8 kHz as compared to normal /ʃ/ spectral energy which ranges from 2-8 kHz (Figure 1(e-f)). The spectral deviation in NAE distorted /ʃ/ is also caused by the coupling of oro-nasal cavity. Therefore, an intensive study of the acoustic characteristics of the NAE distorted fricatives is required for assisting SLPs during speech intervention. Four spectral moments are explored to observed the spectral deviations. The four spectral moments (mean (M1), standard deviation (M2), skewness (M3) and kurtosis (M4)) are computed from the magnitude spectra of the short-term processed fricative signal [16].

4. Temporal variation of spectral characteristics

When the fricative is produced with NAE, the required kinematics of the articulators may be deviated due to forceful compensation to increase the reduced IOP by constricting the VP port. This may change the coordination between the required articulators in the vocal tract and lead to the imprecise production of fricative. This may deviate the temporal variation of the spectral properties (spectral dynamics) of fricative when it is produced with NAE. A state-of-the-art method to analyze the spectral dynamics of the unvoiced sibilant fricative is explored [12, 17]. Here, the spectral dynamics is computed in terms of the variation of peak ERB_N-number, a psychoacoustic measure of the dominant frequency of the spectrum. To compute the peak ERB_N-number, the entire fricative duration is divided into 15 analysis segments of 20 ms duration. The amount of overlap is dependent on the duration of the fricative region. Therefore, the frame-rate varies depending on the speech sample. Then, the DFT magnitude spectrum computed from each segment is passed through a fourth-order gammatone auditory filterbank of 361 filters [17]. Then, the spectral energy at the output of each filter is summed up (termed as "auditory excitation"), and this auditory excitation represents the spectral envelope for a speech frame of a sound unit. The ERB_N-number corresponding to maximum summed energy of the auditory excitation is termed as peak ERB_N-number. The peak ERB_N-number is plotted corresponding to the frame number for the whole fricative duration for both CLP and normal (shown in Figure 2). From the figure, it can be seen that the peak ERB_N-number trajectories for the normal individuals are more straight unlike the case of CLP, where more deviations exist in the individual trajectories for both the fricatives. Also, it can be seen that the mean peak ERB_N-number trajectory of CLP is significantly lower compared to the mean peak of ERB_N-number trajectory of the

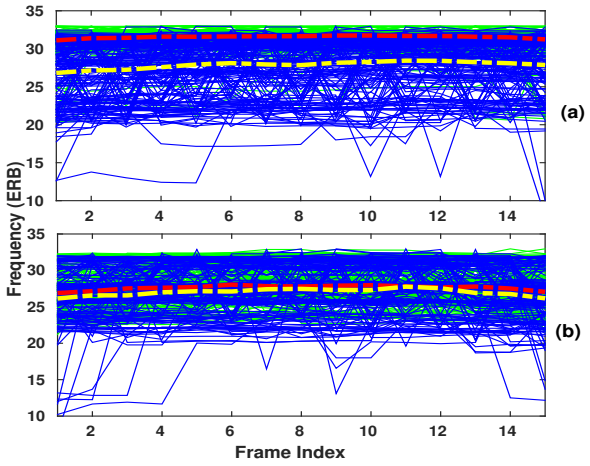

Figure 2: *Illustration of the deviations of peak ERB_N number trajectories of the fricative with NAE (blue) and their mean (yellow dotted) with the trajectories of normal fricative (green) and their mean (red dotted) for (a) /s/ and (b) /ʃ/.*

normal fricative.

Features are derived by characterizing the auditory excitation using DCT. The DCT is used to capture the spectral dynamics present in the auditory excitation. The first 13 dimensions of the DCT coefficients are considered as the compressed representation of the speech frame and termed as auditory excitation features (AEF). However, these 13-dimensional base AEF cannot represent the temporal dynamics of the sounds; therefore, it's derivative (Δ) and acceleration ($\Delta\,\Delta$) features are also computed and augmented with the base AEF. Therefore, the resultant feature dimension is 39 (13 base AEF + 13 Δ + 13 $\Delta\,\Delta$). Unlike the ERB_N study, this feature is computed using an equal frame rate.

5. SVM-based classification

In the current work, an SVM classifier is used for the binary classification of normal fricatives from that of CLP fricatives using a radial basis kernel function (RBF). Separate SVM models are built for each fricative, and for each SVM, the optimum values of the parameters c and γ are experimentally determined using the grid-search method.

For each fricative, four training-testing sets of normal and CLP speech are prepared. Each set in fricative /s/ contains randomly selected 12 normal and 12 CLP children data which are used for training and remaining 4 normal and 4 CLP children data are used for testing. Similarly, each set in fricative /ʃ/ contains randomly selected 11 normals and 11 CLP children data which are used for training and remaining 4 normals and 4 CLP children data are used for testing. Each of the set is assured to be speaker-independent by excluding the same speaker data in the training and testing set at a time. The two-class SVM model is trained and tested for each of the four sets per each fricative. All combinations of the RBF kernel parameters c and γ are considered in the range of $c = [2^{-10}, 2^{-8}, ..., 2^{+8}, 2^{+10}]$ and $\gamma = [2^{-10}, 2^{-8}, ..., 2^{+8}, 2^{+10}]$ during classification. The best accuracy obtained in the considered range of c and γ is reported as a classification result for the specific set of each fricative. Like AEF, Δ and $\Delta\Delta$ variants of spectral moments are also computed. All the spectral moments may not significantly differentiate NAE distorted fricatives from that of normal. Hence,

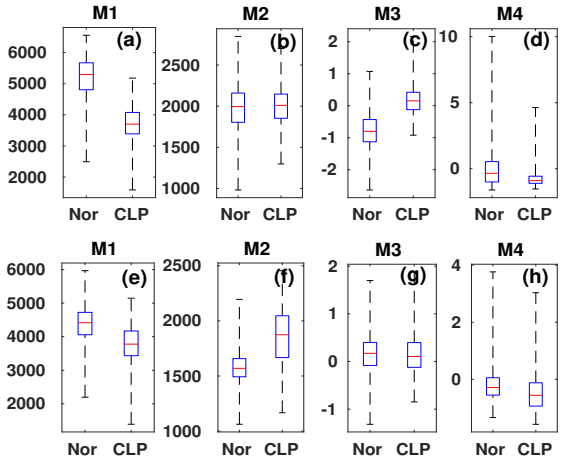

Figure 3: *Box plots of the four spectral moments for (a)-(d) /s/ and (e)-(h) /ʃ/. Nor denotes normal.*

only those features are considered which exhibits statistically significant differences. Further, these features are concatenated with AEF and the resultant feature vector is used to train the SVM model.

6. Results and discussion

The box plots of the four spectral moments are shown in Figure 3 for normal and NAE distorted /s/ and /ʃ/, respectively. Statistical analysis is also performed to observe the discrimination capability of each spectral moment. One-way ANOVA test is conducted for normal and distorted fricatives with the four moments as dependent variables. From the box plots and Table 1 the following observations can be made.

- **Fricative /s/:** The box plots of the four spectral moments for normal and NAE distorted /s/ are shown in Figure 3(a)-(d). The mean and standard deviation of the spectral moments of normal and NAE distorted fricative /s/ is noted in Table 1. The spectral mean (M1) shows that the maximum spectral energy of normal /s/ is concentrated above 5 kHz, whereas NAE distorted /s/ has a maximum spectral density around 3.5 kHz. It is found that for the spectral mean ($p<0.001$) a significant difference is obtained between normal and NAE distorted /s/. The standard deviation (M2) is observed to be slightly higher for NAE distorted /s/ compared to normal /s/, but statistically insignificant in discriminating the two at $p>0.001$. In the case of spectral skewness (M3), NAE distorted /s/ exhibits a more positive spectral slope relative to normal /s/. This conveys that the spectral density in the high-frequency region is relatively low in NAE distorted /s/ as opposed to maximum spectral density concentrated in the higher-frequency region for normal /s/. Statistically, also the skewness is observed to be significantly discriminating normal and NAE distorted /s/. The slight lower peakedness of NAE distorted /s/ indicates that the spectrum is relatively flat compared to normal /s/. Kurtosis (M4) values do not distinguish the normal and NAE distorted /s/ ($p>0.001$).
- **Fricative /ʃ/:** The M1 values of NAE distorted /ʃ/ from Figure 3(e) implies that it is a low frequency dominant signal with spectral energy concentrated around 3 kHz. Whereas, M1 values of normal /ʃ/ is observed 1 kHz

Table 1: *Mean and standard deviation of the spectral moments of normal and NAE distorted fricative /s/ and /ʃ/. Nor denotes normal.*

	M1		M2		M3		M4	
	Nor	CLP	Nor	CLP	Nor	CLP	Nor	CLP
/s/	5189.1 ± 612.34	3690.2 ± 580.22	1970.3 ± 250.85	1988.6 ± 221.20	-0.77 ± 0.50	0.18 ± 0.43	-0.05 ± 1.20	-0.72 ± 0.64
/sh/	4377.9 ± 526.94	3750.7 ± 585.96	1577.0 ± 129.27	1860.6 ± 249.24	0.15 ± 0.37	0.16 ± 0.39	-0.21 ± 0.48	-0.45 ± 0.66

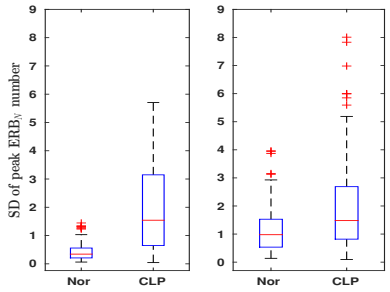

Figure 4: *Box plots of the standard deviation (SD) of peak ERB_N numbers for /s/ (left boxplot), and /ʃ/ (right boxplot).*

higher compared to NAE distorted /ʃ/. The M1 evidently distinguishes the NAE distorted /ʃ/ ($p<0.001$). The M2 value is observed to be higher for NAE distorted /ʃ/ compared to normal /ʃ/ ($p<0.001$). M3 is not distinctively different for normal and NAE distorted /ʃ/, which is observed to be a statistically insignificant measure ($p>0.001$). A slight lower value of M4 is observed for NAE distorted /ʃ/ relative to normal /ʃ/, and it is not significant ($p>0.001$) for the two. Similar observation can be made from Table 1.

The variation of peak ERB_N-numbers for the duration of fricative region is also computed. It can be observed from Figure 4 that the variance of NAE distorted fricative is high compared to the normal case. The possible reason for this high variance may have resulted from the imprecise coordination of the articulators due to reduced intra-oral pressure and the secondary source of turbulence noise in velopharyngeal port. The variance is more in the case of /s/ and less for /ʃ/. A Mann-Whitney U test between the fricatives with NAE and normal fricatives for each feature was performed and a significant difference (< 0.001) is observed.

Table 2: *Performance evaluation of NAE distorted fricative /s/ and /ʃ/ at frame-level.*

Phoneme type	**Accuracy**	**Sensitivity**	**Specificity**
/s/	89.95 ± 2.83	92.70 ± 0.02	87.86 ± 0.07
/ʃ/	94.85 ± 1.43	91.27 ± 0.02	97.09 ± 0.16

Table 3: *Performance evaluation of NAE distorted fricative /s/ and /ʃ/ at utterance-level.*

Phoneme type	**Accuracy**	**Sensitivity**	**Specificity**
/s/	93.27 ± 2.61	95.69 ± 0.01	90.04 ± 0.07
/ʃ/	97.81 ± 1.12	96.88 ± 0.04	98.57 ± 0.02

6.1. Performance evaluation of SVM-based classification

The performance of the SVM-based classification is evaluated at both frame-level and utterance-level. For both cases, overall accuracy, sensitivity, and specificity are considered as the parameter for evaluation. The frame-level results are noted in Table 2. In Table 2, the mean and standard deviation of evaluation parameters at frame level averaged across all the four testing sets are shown for both the fricatives. From the table it can be noted that the overall accuracy of 89.95% is obtained for /s/, whereas, for /ʃ/ the overall accuracy is 94.85%. The improvement in the case of /ʃ/ compared to /s/ may be due to that in our database /ʃ/ is severely distorted due NAE. Similarly, for utterance level overall accuracy, sensitivity, and specificity are noted in Table 3. Out of the total frames in a fricative utterance, if the maximum number of frames corresponds to one particular class, then the respective fricative utterance will be assigned that particular class. Results from Table 3 conveys that fricative /ʃ/ have a overall accuracy of 97.81% compared to 93.27 % for fricative /s/. The sensitivity values are 95.69% and 96.88% for /s/ and /ʃ/, respectively.

7. Summary and future work

This study analyzes the acoustic characteristics of the fricative /s/ produced with NAE in CLP speech. It is found from the study that maximum spectral energy is concentrated in the lower frequency region in case of NAE distorted fricatives. The variance in peak ERB_N-numbers are significantly more in case of fricative with NAE, and it signifies the changes in the peak frequency during fricative region produced with NAE. The variation is more prominent for the fricative /s/ than /ʃ/. An SVM-based classification provides 89.95% and 94.85% overall accuracy in frame-level for /s/ and /ʃ/, respectively. Whereas, at the utterance level the accuracies are 93.27% and 97.81% for /s/ and /ʃ/, respectively. This is a preliminary work, which is done on a small database. Future work is planned to perform the experiment on a relatively large database. The acoustic characterization of the other pressure consonants (stops and affricates) distorted by NAE is also planned. It is also necessary to study the different levels of NAE, mild or severe, which may provide more information about the acoustics characteristics of NAE.

8. Acknowledgement

The authors would like to thank Prof. M. Pushpavathi and Prof. Ajish Abraham, AIISH, Mysore, for providing CLP speech samples and insight about CLP speech disorder. 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.

9. References

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

[2] B. J. Costello, R. L. Ruiz, and T. A. Turvey, "Velopharyngeal insufficiency in patients with cleft palate," *Oral and Maxillofacial Surgery Clinics of North America*, vol. 14, no. 4, pp. 539–551, November 2002.

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

[4] K. Bzoch, "Introduction to the study of communicative disorders in cleft palate and related craniofacial anomalies," *Communicative disorders related to cleft lip and palate. 5th ed. Austin: pro-ed*, pp. 3–66, 2004.

[5] D. W. Warren, "Conpensatory speech behaviors in individuals with cleft palate: A regualtion/control phenomenon?" *Cleft Palate Journal*, vol. 23, no. 4, pp. 251–260, October 1986.

[6] A. M. A. Ali, J. V. der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," *J. Acoust. Soc. Am. 109 (5), Pt. 1, May 2001*, vol. 109, no. 5, pp. 2217–2235, May 2001.

[7] A. L. Baylis, B. Munson, and K. T. Moller, "Perceptions of audible nasal emission in speakers with cleft palate: A comparative study of listener judgments," *Cleft PalateCraniofacial Journal*, vol. 48, no. 4, July 2011.

[8] M. J. Cler, Y.-A. S. Lien, M. N. B. andTalia Mittelman, K. Downing, and C. E. Stepp, "Objective measure of nasal air emission using nasal accelerometry," *Journal of Speech, Language, and Hearing Research*, vol. 59, pp. 1018–1024, October 2016.

[9] H. Dotevall, A. Lohmander-Agerskov, , H. Ejnell, and B. Bake, "Perceptual evaluation of speech and velopharyngeal function in children with and without cleft palate and the relationship to nasal airflow patterns," *Cleft Palate Craniofac J.*, vol. 39, pp. 409–424, 2002.

[10] K. Bettens, F. L. Wuyts, and K. M. Van Lierde, "Instrumental assessment of velopharyngeal function and resonance: A review," *Journal of communication disorders*, vol. 52, pp. 170–183, 2014.

[11] D. J. Zajac, R. Mayo, R. Kataoka, and J. Y. Kuo, "Aerodynamic and acoustic characteristics of a speaker with turbulent nasal emission: A case report," *The Cleft palate-craniofacial journal*, vol. 33, no. 5, pp. 440–444, 1996.

[12] P. F. Reidy, "The spectral dynamics of voiceless sibilant fricatives in english and japanese," Ph.D. dissertation, The Ohio State University, 2015.

[13] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of english fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[14] K. Nataraj, P. C. Pandey, and H. Dasgupta, "Estimation of place of articulation of fricatives from spectral characteristics for speech training," *Proc. Interspeech 2017*, pp. 339–343, 2017.

[15] K. Iskarous, C. H. Shadle, and M. I. Proctor, "Articulatory–acoustic kinematics: The production of american english/s/," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 944–954, 2011.

[16] Y. Feng, G. J. Hao, S. A. Xue, and L. Max, "Detecting anticipatory effects in speech articulation by means of spectral coefficient analyses," *Speech communication*, vol. 53, no. 6, pp. 842–854, 2011.

[17] P. Reidy, "Spectral dynamics of sibilant fricatives are contrastive and language specific," *J. Acoust. Soc. Am.*, vol. 140, no. 4, pp. 2518–2529, October 2016.