IMPACT OF SPEAKING RATE ON THE SOURCE FILTER INTERACTION IN SPEECH: A STUDY

Tilak Purohit, Achuth Rao MV, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bengaluru 560012, India

ABSTRACT

Source filter interaction (SFI) explains the drop in pitch caused due to the constriction in the vocal tract during voiced consonant production in a vowel-consonant-vowel (VCV) sequence. In this work, we examine how the drop in pitch alters when such a VCV sequence is spoken at three different speaking rates - slow, normal and fast. In the absence of electroglottograph (EGG) recording, a high resolution pitch contour is determined using a glottal closure instant (GCI) detector. For this, in this work, firstly, five different GCI detector and pitch estimation techniques are compared against EGG based pitch estimates on a small dataset where simultaneous EGG recordings are available. Yet Another GCI Algorithm (YAGA) is found to be the best choice among all. For examining the impact of speaking rate on SFI, VCV recordings from six subjects with five vowels (/a/, /e/, /i/, /o/, /u/) and five consonants (/b/, /d/, /g/, /v/, /z/) at three speaking rates are used. The study reveals a significant difference in the pitch drop values between slow and fast rates, with increasing pitch drop as speaking rate reduces. For slow speaking rate, vowel /o/ and /u/ tend to show higher pitch drop values compared to remaining vowels.

***Index Terms*—** Source filter interaction, Pitch drop, Speaking rate

1. INTRODUCTION

In human speech production, the characteristics of glottal source are often affected by the vocal tract, this is known as the source filter interaction (SFI). The interaction between the glottal source and vocal tract has been investigated by several researchers in the past [1–3]. Titze and Palaparthi [4] broadly divide the SFI into two levels. Level-1 describes the effect of the vocal tract on the glottal flow features, whereas level-2 describes the effect of the vocal tract on the pitch[1]. In this study, we explore the level-2 SFI, where there is an involuntary change in the pitch due to vocal tract configurations. The involuntary changes in the glottal vibration occurs during the changes in the "intrinsic pitch" of some high vowels [5, 6]. This could be caused either by the coupling of the glottis and the vocal tract [5–7] or could be due to the effect of the tongue-pull [6]. Vowel formants were also used to examine the effect of coupling between the oral and the sub-glottal cavities [8]. Stevens [9] carried out the studies on SFI for the speech sounds, such as fricatives and stops. It has been shown that, constriction along the vocal tract during the consonant production in a VCV utterance causes a drop in pitch. The amount of pitch change depends on the degree and location of the constriction [10].

This study focuses on the percentage of pitch change in the voiced consonant region compared to the vowel region in a vowl-consonant-vowel (VCV) sequence. The study also quantifies the percentage of pitch change in voiced consonant in a VCV sequence across three different speaking rates namely, slow, normal and fast. The study aims to provide a scientific understanding and document the SFI happening during human speech production. Typically, the SFI study uses electroglotogram (EGG) signals [11] to estimate the pitch. But the collection of data with parallel EGG and acoustic signal can be time consuming and expensive. Hence, we propose to use a glottal closure instant (GCI) detection algorithm to estimate pitch. Mittal et al. [10] used zero-frequency filtering (ZFF) [12] method to extract features of glottal source excitation directly from the speech signals for the SFI analysis. Recently, there are several GCI detection methods proposed in the literature. To determine the best algorithm to quantify the pitch drop, we first compare five-pitch estimation methods with the pitch estimated from the EGG signal using a corpus which has both acoustics and EGG signals. We compare four GCI based pitch estimation techniques, namely, Yet Another GCI Algorithm (YAGA) [13], glottal closure/opening instant estimation forward-backward algorithm (GEFBA) [14], ZFF [12], and Dynamic Programming Phase Slope Algorithm (DYPSA) [15], and one pitch contour estimation technique without GCI detector, namely, sawtooth waveform inspired pitch estimator (SWIPE) [16]. We find that the pitch drop in the voiced consonant region estimated from the YAGA is closest to that estimated from the EGG signals. Hence, we use the YAGA for all experiments in this study on the impact of speaking rate on SFI.

Most work on SFI focus on the VCV (or CV) sequence with the vowel being /a/ spoken at a normal speaking rate (SR). In this work, we explore the SFI in the context of different SRs. We hypothesize that the pitch drop during a voiced consonant in the context of the different vowels could be different given that the articulatory shapes for different vowels are different. The speaking rate affects the movement of the articulators and can have a significant effect on the pitch drop as well. Hence, we measure the pitch drop during voiced consonant in the VCV sequence using a dataset containing six subjects speaking VCV sequences with five vowels (/a/, /e/, /i/, /o/, and /u/) and five consonants (/b/, /d/, /g/, /v/, and /z/) at three speaking rates, namely, slow, normal and fast. We find that the pitch drop increases as the SR reduces with significant difference between slow and fast speaking rates for all 25 VCV combinations except for /a/b/a/, /e/g/e/ and /i/b/i/. The pitch drop is found to depend significantly on the vowels at slow speaking rate.

2. DATASET

VCV sequence is typically used for SFI study as it is relatively easy to distinguish the vowels and consonants regions for the analysis if the vowel is present on both sides of the consonant [10]. In this preliminary study of SFI, there are two parts, first is to find out which GCI detection algorithm is the best to estimate the pitch from the acoustic signal particularly in the context of SFI study, and second, to investigate the pitch drop trend across different vowels and consonant combinations in three different speaking rates. Considering these factors, two different datasets are considered for this study. While both datasets have clean speech acoustic recordings of VCV

[1] By 'pitch' in this work, we refer to the fundamental frequency

	/b/	/d/	/g/	/v/	/z/
/a/	0.12(0.02)	0.10(0.03)	0.10(0.03)	0.14(0.07)	0.16 (0.06)
	0.09(0.01)	0.05(0.01)	0.06(0.01)	0.06(0.01)	0.07 (0.01)
	0.05(0.02)	0.04(0.01)	0.04(0.01)	0.04(0.01)	0.05 (0.01)
/e/	0.13(0.03)	0.11(0.03)	0.12(0.04)	0.17(0.08)	0.18(0.07)
	0.08(0.01)	0.08(0.02)	0.08(0.03)	0.08(0.03)	0.10(0.03)
	0.05(0.01)	0.04(0.01)	0.04(0.01)	0.04(0.01)	0.06(0.01)
/i/	0.14(0.05)	0.14(0.09)	0.16(0.12)	0.18(0.11)	0.19(0.11)
	0.09(0.02)	0.08(0.02)	0.10(0.04)	0.07(0.03)	0.10(0.04)
	0.06(0.01)	0.06(0.01)	0.04(0.02)	0.04(0.01)	0.06(0.01)
/o/	0.18(0.10)	0.17(0.12)	0.20(0.15)	0.17(0.10)	0.19(0.10)
	0.10(0.02)	0.09(0.05)	0.10(0.04)	0.10(0.03)	0.12(0.05)
	0.07(0.01)	0.05(0.02)	0.06(0.02)	0.06(0.01)	0.06(0.01)
/u/	0.13(0.04)	0.14(0.07)	0.16(0.09)	0.18(0.09)	0.16(0.05)
	0.08(0.01)	0.07(0.01)	0.10(0.02)	0.10(0.03)	0.09(0.02)
	0.05(0.01)	0.05(0.01)	0.06(0.02)	0.07(0.02)	0.06(0.01)

Table 1. The average duration of the consonant region (in seconds) for the VCV combinations in the SPIRE VCV corpus. The 3 rows in a cell (from top to bottom) corresponds to slow, normal and fast speaking rates, in blue, green and magenta color respectively. $(\cdot)$ indicates the standard deviation.

stimuli, the first corpus (EGG corpus) has EGG signal recordings in parallel to the acoustic recordings at normal speaking rate while the second corpus (SPIRE VCV corpus) has only acoustic speech recordings (without any EGG recordings) in three different speaking rates namely, slow, normal, and fast. Details of these two corpora are described in the subsections below.

2.1. EGG corpus

The EGG corpus consists of isolated VCV samples for five voiced consonants, namely /b/, /g/, /j/, /v/, /z/ and vowel /a/. Each VCV was recorded 7 times each from nine subjects, three females and six males with their average age being 23 and 25 years, respectively. None of the subjects were reported to have any speech disorder. A total of 315 samples (9 subjects $\times$ 5 VCVs $\times$ 7 repetitions) are present in this corpus. The data collection was carried out in a soundproof room in SPIRE Lab at IISc. Simultaneous recordings of the speech and the EGG signal were obtained for each VCV. Acoustic recording was done using a Sennheizer e822S microphone (Wedemark, Germany) and EGG signal was obtained using VoceVista(Roden, The Netherlands). The audio and EGG were recorded at 16 kHz. The begin and end time-stamps of each VCV recording and the C-boundaries within every VCV were manually marked by listening as well as examining the waveform and the spectrogram.

2.2. SPIRE VCV corpus

The SPIRE VCV dataset consists of utterance of the type "speak VCV today", having all combinations of five consonants (C) namely /b/, /d/, /g/, /v/, /z/ and five vowels (V) /a/, /e/, /i/, /o/ and /u/, vowels being on both the sides of the consonant. Each utterance was collected in three different speaking rates and each VCV sample had 3 repetitions. The data was recorded from six subjects, three females and three males of the age range 18-22years. Thus, we have a total of 1350 (=5 consonants $\times$ 3 repetitions$\times$ 5 vowels $\times$ 3 rates$\times$ 6 subjects) recordings.

To make sure that each subject controls the speaking rate uniformly while speaking, a demo session was taken before the actual recording, where the speaking rate modulation was practiced. Each subject was instructed to speech normally in normal speaking rate. In slow (fast) speaking rate, every subject was asked to speak twice as slow (fast) as the normal speaking. The average duration of the consonant region for all 450 samples in slow speaking rate is found to be 0.15 ± 0.08 seconds. The same for normal and fast rates are 0.08 ± 0.03 seconds, and 0.05 ± 0.01 seconds. These duration values of consonants across rates suggest that subjects could follow the given instructions well during recording. Table 1 presents the duration of consonant for each consonant vowel combinations in three speaking rates. All recordings were done at the SPIRE Lab's soundproof studio, Indian Institute Science, Bangalore, India. The VCV boundaries were manually annotated. The boundaries were marked by observing the spectrogram, the raw waveform, and the glottal pulses (obtained using Praat [17]) simultaneously using an in-house built MATLAB based annotation tool.

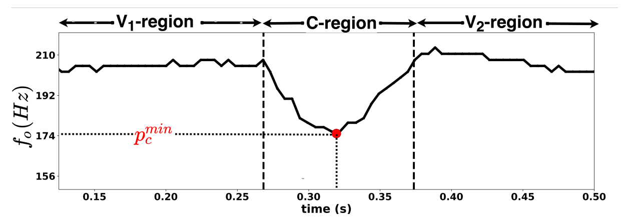

Fig. 1. The black curve shows the pitch contour trajectory, and the vertical dotted lines mark the consonant (C) boundaries.

3. PITCH DROP MEASURE IN THE PROPOSED SFI STUDY

In this work, we compute the measure of SFI using two steps. In the first step, we compute the pitch contour from the acoustic signal using a pitch estimation technique. An illustration of pitch contour for a sample VCV is shown in Fig. 1, where V_1 and V_2 are the two identical vowels before and after the consonant. Following this, we compute a measure of SFI as the percentage pitch drop in the C-region compared to the V_1-region as follows: $p_\delta = \frac{(p_{V1}^{med} - p_C^{min}) \times 100}{p_{V1}^{med}}$, where p_{V1}^{med} is the median of the pitch in the last $2/3^{rd}$ of the V_1-region and p_C^{min} is the minimum pitch in the C-region as shown in Fig 1.

4. SELECTION OF THE GCI DETECTOR IN THE PROPOSED STUDY

4.1. Data preparation

The best GCI detection scheme was determined on the EGG corpus using four GCI detection and one pitch estimation algorithms. The EGG was further refined by removing the cases (referred to as outliers) where the estimated pitch from different algorithms had errors on visual inspection. Outliers also included samples where the voicing signature was not observed in the consonant region of the VCV sequence. The absence of the voicing signature makes it difficult for the waveform-based GCI detection algorithms to locate the GCIs accurately. An example of such a sample is depicted in Fig. 2(C). A total of 98 outliers were removed from the corpus. The remaining 217 samples from the EGG corpus were considered for the selection of best GCI detection algorithm.

4.2. Selection of the best GCI detector

Firstly, five distinct pitch contours were computed, out of which four were derived from the GCIs using GEFBA, YAGA, ZFF, DYPSA

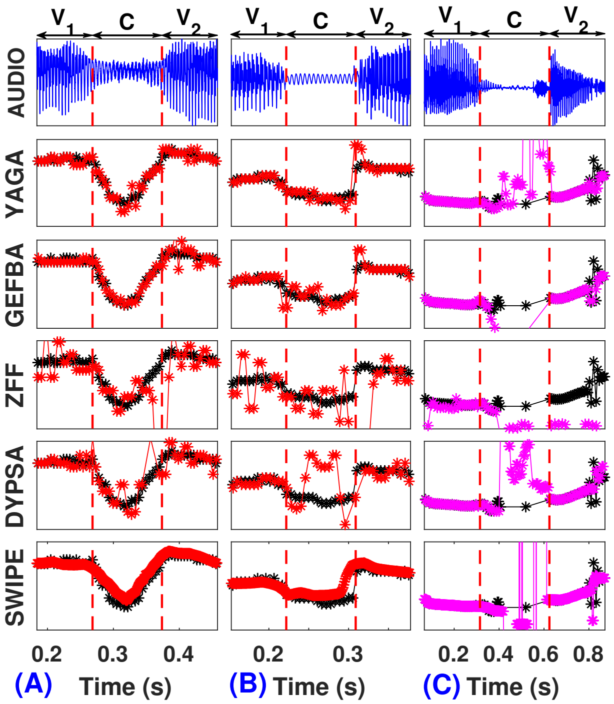

Fig. 2. Starred black trajectories depict the ground truth SIGMA derived pitch contours, the superimposed red and the magenta trajectories (starred) show the pitch contours derived via various pitch estimation schemes indicated on the Y axis of column (A) subplots. Red dashed horizontal lines are the consonant boundaries. The plot (A) and (B) show the efficacy of the YAGA over other schemes. Plot(C) shows a typical outlier where voicing signature in C region is missing making it difficult for the pitch estimating algorithms to locate the GCIs.

and the remaining one was estimated from SWIPE, which directly estimates the pitch contour from the speech waveform. SIGMA algorithm was used to obtain pitch contour from the recorded EGG signal. Illustrative examples of these are shown in Fig 2. The implementations of SIGMA and the DYPSA algorithms were taken from the VOICEBOX Toolbox [18]. For other algorithms such as GEFBA, YAGA, ZFF, and SWIPE, the implementations were provided by their respective authors. The derived pitch contour was then passed through a median filter with a window size of 3 samples, to remove the noise in the pitch trajectory. A smaller window size (3 samples) was preferred so as not to smooth out the dip in pitch contour in the voiced consonant region. Then we compute the measure of source-filter interaction (SFI), i.e., p_δ. p_δ computed on the pitch contour from the SIGMA algorithm was considered as the ground truth as the GCIs obtained were based on the EGG signals. Finally, the algorithm which gave the value of p_δ closest to that from the SIGMA scheme was determined to be the best scheme.

Table 2 shows the average p_δ along with their standard deviation (SD) in bracket obtained using SIGMA and all pitch estimation algorithms. The last column shows the p-value from t-test to examine if there is a significant difference between the pitch drop as observed in the case of SIGMA and each of the algorithms considered. Considering the average p_δ values, it is clear that the YAGA algorithm performs closest to the ground truth p_δ among all algorithms considered. This is also clear from Fig 2(A) & (B), which show the pitch trajectories estimated using five pitch estimation schemes. This result is similar to the finding of [19], where YAGA performs better

	p_δ (EGG Corpus)	p-value
SIGMA	13.42 (11.98)	-
YAGA	12.28 (6.79)	0.30
GEFBA	12.06 (7.71)	0.25
ZFF	11.98 (7.92)	0.23
DYPSA	16.20 (9.96)	0.04
SWIPE	9.86 (9.62)	0.004

Table 2. The average of p_δ across all 217 considered samples in EGG corpus, $(\cdot)$ indicates the standard deviation. Last column shows p-value from t-test with null hypothesis that p_δ from SIGMA and a pitch estimation algorithm have identical mean.

with clean speech data, as it considers the glottal flow derivatives obtained by inverse filtering the speech signal. The difference between the average p_δ values from SIGMA and YAGA for different VCVs are follows: /a/b/a/ 0.86; /a/g/a/ 5.9; /a/j/a/ 3.02; /a/v/a/ -0.43 and /a/z/a/ -2.82. These values show that YAGA performed the best for consonant /b/ and /v/ followed by /z/, /j/, and /g/. It is interesting to observe that YAGA was able to perform well for the consonant sounds with labial constriction. Its performance was found to be the worst in case of /g/, which has a velar constriction.

5. IMPACT OF RATE ON SFI

5.1. Data preparation

In a manner similar to that for the EGG corpus, the outliers (as depicted in Fig. 2(C)) were removed from the SPIRE VCV corpus by visually inspecting 1350 VCV samples. A total of 220 samples were detected as outliers (13.6% from slow, 15.5% from normal and 19.8% from fast) and were removed, thus retaining 83.7% of the samples from the dataset, for the study on impact of rate on SFI. The number of samples considered for the study for each vowel and consonant combination is shown in Table 3. YAGA algorithm was used for the pitch estimation and the pitch contours were passed through a median filter with window size of 3 samples, to reduce the noise in the contour trajectory.

We compute p_δ for all VCVs with three speaking rates, from the SPIRE VCV corpus. We study the variation in p_δ values for different speaking rates. We perform a t-test between p_δ in slow and fast speaking rate samples [20]. This is also done between normal and fast speaking rates. We also perform the analysis of variance (ANOVA) for different vowels and consonants with a given speaking rate [20].

	/b/	**/d/**	**/g/**	**/v/**	**/z/**
/a/	45	46	33	49	40
/e/	43	47	46	49	42
/i/	50	41	47	47	47
/o/	41	45	46	42	43
/u/	50	46	53	47	45

Table 3. Sample distribution in the SPIRE VCV corpus after the outlier removal; 54 is maximum possible value a cell could have (6 subjects $\times$ 3 repetitions $\times$ 3 rates.)

5.2. Results and discussion

Fig. 3 shows the p_δ for all VCVs and different speaking rates. It is clear from the figure that the mean and median value of p_δ increase as the speaking rate reduces. It could be because, in case of a slow

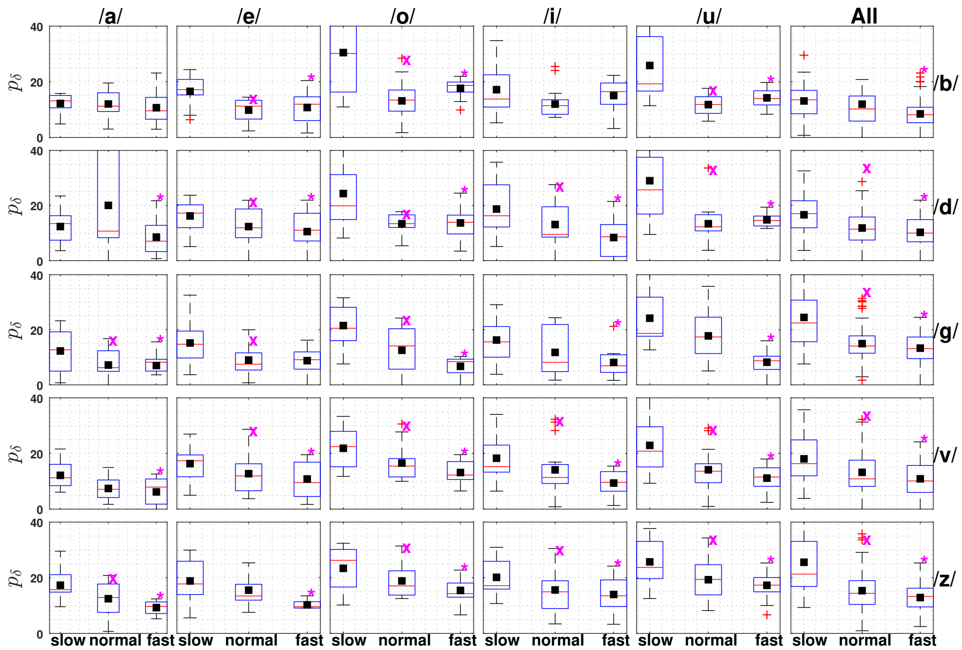

Fig. 3. The box plot comparison of p_δ for different VCVs. The last column shows the pitch drop values when all vowels are combined. The red line shows the median, black square shows the mean, blue box is the 75% confidence interval and $+$ shows the outliers. The $*$ indicates the cases with significant difference between the fast and slow rates. The $\times$ indicates the cases with significant difference between fast and normal speaking rates.

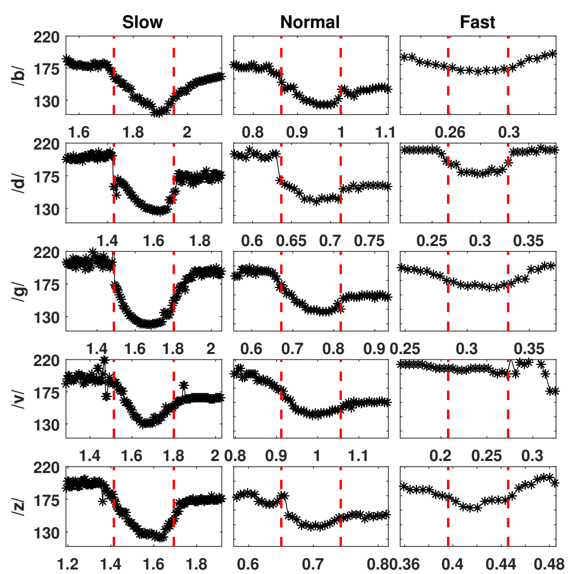

Fig. 4. Illustration of pitch trajectory for different consonants with vowel /u/ at three different speaking rates. The consonant boundary is marked with red dotted line.

speaking rate, the duration of constriction is longer causing a larger effect on the vocal fold vibration. The statistical test shows that there is a significant difference (p <0.05) between p_δ in slow and fast speaking rates in all 25 VCVs except in the cases of /a/b/a/, /e/g/e/ and /i/b/i/. In most VCVs, there is a significant difference (p <0.05) between fast and normal speaking rates as well. Fig. 4 presents a comparison of exemplary pitch contours for different consonants with vowel /u/ at three different speaking rates. It is clear from the figure that the pitch drop in the voiced consonant region reduces as speaking rate increases. It is also clear from Fig. 3 that when samples from all vowels are combined, p_δ shows a trend similar to an individual vowel, and there is a significant difference between p_δ from slow vs. fast and normal vs. fast speaking rates except in the case of /b/ between normal and fast speaking rates. ANOVA test reveals that the pitch drops for different consonants within one vowel for slow rate are significantly different. When all vowels are combined (last column in Fig. 3), the variance of p_δ reduces in the fast/normal rate compared to the slow rate. The pitch drops for different vowels within one consonant in slow speed are not found to be significantly (p-value>0.1) different, but for the normal speed, the difference is statistically significant (p-value<.02) in all cases except for ‘V’. It could be because, for slow speed, the constriction reaches a steady-state in the consonant region irrespective of the vowels. As the speed increases, the degree of constriction varies depending on the vowel, and, hence, there is a significant difference in p_δ across vowels.

6. CONCLUSION

In this study, we quantify SFI as the percentage of pitch drop in a voiced consonant compared to the vowel region in a VCV sequence. First, we use a corpus with parallel acoustic and EGG recordings to determine YAGA as the best GCI detector cum pitch estimation scheme for studying SFI. We further estimate the pitch drop using YAGA on a larger corpus having VCV sequences spoken by six subjects with five vowels (/a/, /e/, /i/, /o/, /u/) and five consonants (/b/, /d/, /g/, /v/, /z/) at three speaking rates, slow, normal, and fast. When we examine the impact of speaking rate on SFI on this larger corpus, it reveals a significant difference in the pitch drop values between slow and fast rates, with increasing pitch drop as the speaking rate reduces. For slow speaking rate, all consonants with vowels /o/ and /u/ tend to show higher pitch drop values than remaining vowels. In future, we plan to extend this study with larger consonant set under asymmetric VCV stimuli with different vowels.

7. REFERENCES

[1] Jorge C. Lucero, Kélem G. Lourenço, Nicolas Hermant, Annemie Van Hirtum, and Xavier Pelorson, "Effect of source–tract acoustical coupling on the oscillation onset of the vocal folds," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 403–411, 2012.

[2] Ingo Titze, Tobias Riede, and Peter Popolo, "Nonlinear source-filter coupling in phonation: Vocal exercises," *The Journal of the Acoustical Society of America*, vol. 123, pp. 1902–15, 05 2008.

[3] Ingo R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2733–2749, 2008.

[4] Ingo Titze and Anil Palaparthi, "Sensitivity of source-filter interaction to specific vocal tract shapes," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 2507–2515, 10 2016.

[5] William G. Ewan and John J. Ohala, "Can intrinsic vowel f0 be explained by source/tract coupling?," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 358–362, 1979.

[6] John Ohala and Brian Eukel, *Explaining the Intrinsic Pitch of Vowels*, vol. 60, pp. 207–215, 01 1987.

[7] E. Vilkman, U.K. Laine, and J. Koljonen, "Supraglottal acoustics and vowel intrinsic fundamental frequency; an experimental study," *Speech Communication*, vol. 10, no. 4, pp. 325 – 334, 1991.

[8] Xuemin Chi and Morgan Sonderegger, "Subglottal coupling and its influence on vowel formants," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1735–1745, 2007.

[9] Kenneth N. Stevens, "Airflow and turbulence noise for fricative and stop consonants: Static considerations," *The Journal of the Acoustical Society of America*, vol. 50, no. 4B, pp. 1180–1192, 1971.

[10] Vinay Kumar Mittal, B. Yegnanarayana, and Peri Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1932–1941, 2014.

[11] Achuth Rao M V, Shiny Victory J, and Prasanta Kumar Ghosh, "Effect of source filter interaction on isolated vowel-consonant-vowel perception," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. EL95–EL99, 2018.

[12] 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.

[13] Mark RP Thomas, Jon Gudnason, and Patrick A Naylor, "Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2011.

[14] Andreas I Koutrouvelis, George P Kafentzis, Nikolay D Gaubitch, and Richard Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 316–328, 2015.

[15] Patrick A Naylor, Anastasis Kounoudes, Jon Gudnason, and Mike Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2006.

[16] Arturo Camacho and John G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[17] Boersma Paul and Weenink David, "Praat: doing phonetics by computer," .

[18] Mike Brookes et al., "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, vol. 47, 1997.

[19] A. R. MV and P. K. Ghosh, "PSFM—a probabilistic source filter model for noise robust glottal closure instant detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1645–1657, 2018.

[20] Cyril Harold Goulden, "Methods of statistical analysis.," 1939.