AUGUST 06 2018

Effect of source filter interaction on isolated vowel-consonant-vowel perception

Achuth Rao M V; Shiny Victory J; Prasanta Kumar Ghosh

J Acoust Soc Am 144, EL95–EL99 (2018)
https://doi.org/10.1121/1.5049510

Effect of source filter interaction on isolated vowel-consonant-vowel perception

Achuth Rao M V,[1,a)] Shiny Victory J,[2] and Prasanta Kumar Ghosh[1]
[1]*Indian Institute of Science, Bangalore, India*
[2]*Mepco Schlenk Engineering College, Sivakasi, India*
achuthr@iisc.ac.in, jshinyv96@gmail.com, prasantg@iisc.ac.in

Abstract: Source-filter interaction explains the drop in pitch in voiced consonant due to constriction in the vocal tract during vowel-consonant-vowel (VCV) production. In this work, a perceptual study is conducted where the pitch contour in the voiced consonant region is modified to four different levels and a listening test is performed to assess the naturalness of the VCVs synthesized with the modified pitch contour. The listening test with 30 listeners shows no statistically significant difference between the naturalness of the original and synthesized VCVs with modified pitch indicating that pitch drop due to source-filter interaction may not be critical for the perceived naturalness of VCVs.

[DDO]
Date Received: May 9, 2018 **Date Accepted:** July 18, 2018

1. Introduction

In human speech production, the time varying vocal tract is excited by the airflow from the glottis. The speech sounds, thus produced, can be classified broadly into two classes: (1) the unvoiced sounds where the vocal fold remains open, and (2) the voiced sounds where the vocal fold vibrates in a quasi-periodic manner (Titze and Alipour, 2006). Pitch is the perceptual counterpart of this quasi-periodicity in the acoustic signal. The acoustics of human speech production is typically represented by a linear source-filter (SF) model. This linear model is based on the assumption that the source is independent of the filter (Fant, 1971), where the source signal is convolved with the filter impulse response to generate speech.

The linear SF model, while simple, often does not hold good for representing speech. This is because, in human speech production, glottal source characteristics are often affected by those of the vocal tract which is known as the source-filter interaction (SFI). The interaction between the glottal source and the vocal tract system has been studied by several researchers (Lucero *et al.*, 2012; Titze *et al.*, 2008; Titze, 2008). The SFI is broadly divided into two levels (Titze and Palaparthi, 2016). Level-I SFI describes the effect of vocal tract on the glottal flow features such as skewness and the delay of peak glottal airflow relative to the peak glottal area. Level-II SFI describes the vocal tract effect on the pitch. In this study, we focus on the level-II SFI, where the pitch changes involuntarily because of the vocal tract shape. The involuntary changes in the glottal vibration occur during changes in the "intrinsic pitch" of some high vowels (Ewan and Ohala, 1979; Ohala and Eukel, 1987). The effect could be due to either coupling between the vocal tract and the glottis (Ewan and Ohala, 1979; Ohala and Eukel, 1987; Vilkman *et al.*, 1991) or the tongue-pull effect (Ohala and Eukel, 1987). The effects of coupling between oral and sub-glottal cavities were examined through vowel formants (Chi and Sonderegger, 2007). Studies on the SFI were also carried out for other categories of speech sounds, such as fricatives and stops (Stevens, 1971). It has also been shown that the pitch decreases in the voiced consonant region in a vowel-consonant-vowel (VCV) because of the constriction along the vocal tract during the consonant production. The amount of pitch change depends on the degree and location of the constriction (Mittal *et al.*, 2014). The perceptual importance of the different interaction effects has not been very comprehensively examined. However, Fant and Liljenkrants (1979) studied the perceptual significance of the formant decay rate. Nord *et al.* (1984) studied four acoustic effects of interaction: skewing, truncation, dispersion, and superposition. They showed that for complex sounds there is no perceptual discrimination with or without SFI in terms of those four

[a)]Author to whom correspondence should be addressed.

acoustic effects. Båvegård (2009) studied the perceptual significance of the effect of glottal ripple due to SFI.

The linear SF model has been widely used in text to speech synthesis (TTS) systems (Taylor, 2009), where the goal is to synthesize natural sounding speech. In a TTS system, the vocal tract filter is represented by either linear prediction coefficients or Mel-cepstral coefficients (Yoshimura *et al.*, 1999) and the glottal source is represented by either white noise or periodic pulses with a period identical to the pitch period (Fant, 1971). The SF model, while attempting to mimic the human speech production system, does not explicitly model the SFI. While it is known that level-II SFI affects the glottal source vibration in speech production, it is not clear how a SF model, that does not model the SFI accurately, affects speech perception. A study that quantifies the effect of SFI on speech perception could provide insight for explicitly modeling SFI in a TTS system. While there have been a number of studies on the perceptual significance of various effects of SFI, to the best of our knowledge, there is no study that quantifies the perceptual significance of the SFI that affects the glottal vibration.

In this current study, we examine the role of involuntary pitch change due to level-II SFI on the naturalness of speech using a listening test. To avoid the voluntary pitch changes because of the text being spoken, we focus our study only on the isolated VCVs. Figure 1 shows a sample pitch contour for the VCV /aba/. A drop in the pitch value in the consonant region is clearly observed compared to the pitch in the surrounding vowel regions. In order to understand the effect of SFI on speech perception, we modify the pitch contour within the consonant region without modifying the pitch in the vowel regions while maintaining an overall smooth pitch contour. We quantify the effect of such pitch modification on the naturalness of the synthesized speech compared to the original speech using a listening test. This is carried out in two steps. At first, we separate the filter and the source signal including the pitch contour using a WORLD vocoder (Morise, 2016; Morise *et al.*, 2016). In the second step, we propose an algorithm that modifies the dip in the pitch contour only in the consonant region (because of SFI) while maintaining the continuity of the pitch contour. Finally, the modified pitch contour, along with the original filter response, is used to synthesize the speech signal. The synthesized speech signals corresponding to different modified pitch levels are compared against the original speech using a listening test with thirty listeners using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test (ITU, 2003). Results show that there is no significant difference between the naturalness of the original and synthesized speech signals. The experimental findings suggest that the drop in pitch due to SFI in VCV production may not be critical for perceiving its naturalness.

2. Dataset

In human speech production, speech sounds are produced as part of one or more syllables such as CV, VCV, or VCCV, consisting of vowels (V) and consonants (C) (Stevens, 2000). If the vowel is present on both sides of the consonant, then it is relatively easy to distinguish the vowel and consonant regions for analysis (Mittal *et al.*, 2014). In this study, isolated VCV voice samples were recorded for five voiced consonants: (1) /b/, (2) /g/, (3) /v/, (4) /z/, and (5) /j/. All VCVs have vowel /a/ on either side of the voiced consonant. Each of the VCVs was recorded five times from 11 subjects (5 females and 6 males). The average age of the female and male subjects was 23(±6)

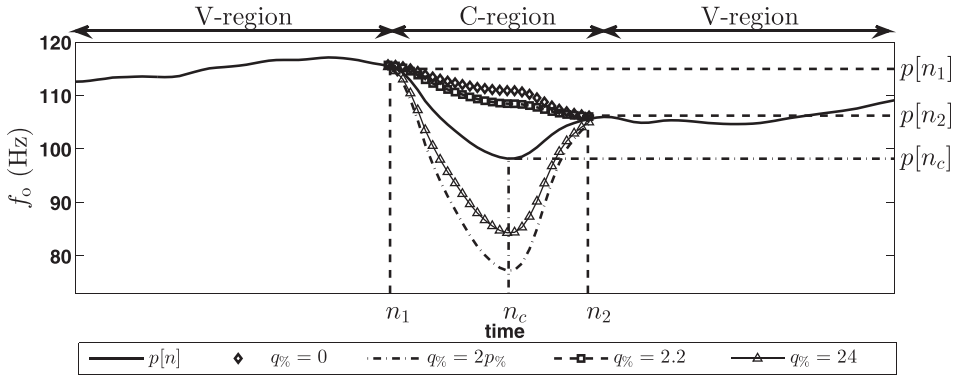

Fig. 1. Comparison of the original pitch contour and the modified pitch contours with $q_\% = 0$, $2p_\%$, 2.2, 24 for /aba/. The boundaries of the consonant and vowel regions are marked with dashed vertical lines.

and 25(±6) years, respectively. Subjects were graduate students at the institute. None of the subjects were reported to have any speech disorders. Thus, a total of 275 (=11 subjects × 5 VCV × 5 repetitions) VCV samples were recorded for the study.

The data collection was carried out in a soundproof room. Simultaneous recordings of the speech and the Electroglottograph (EGG) signal were obtained for each VCV using a Sennheizer e822S microphone (Wedemark, Germany) and EGG from VoceVista (Roden, The Netherlands). The audio and EGG were recorded at 16 kHz with 16 bits/sample. The microphone was kept a distance of ∼12 cm from the subject's mouth. The sound pressure level at the soundproof room was 30 dBA. 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.

3. Proposed study of the effect of SFI on VCV perception

The proposed study on the effect of SFI on VCV perception is carried out in two steps. The first step is to separate the filter and the source signal including the pitch contour using the WORLD vocoder given a VCV recording. In the second step, we modify the pitch contour in the C-region using a proposed algorithm and synthesize the speech using filter parameters obtained in the first step and the modified pitch contour. Details of these steps are described below.

3.1 Speech analysis

A given VCV sample signal s is decomposed into source and filter parameters following a SF model. In order to study the effect of pitch modification on the naturalness of the synthesized speech, it is required that the chosen SF model does not incur any loss in naturalness due to its analysis and synthesis steps. STRAIGHT (Kawahara *et al.*, 2008) and WORLD (Morise *et al.*, 2016) vocoders are well known for this purpose. However, we use the WORLD vocoder, as it is superior to the STRAIGHT vocoder (Morise *et al.*, 2016). The WORLD vocoder decomposes the signal into spectral envelope, aperiodicity, and pitch p for every frame comprising N_s samples (Morise *et al.*, 2016). The quality of this decomposition depends on the accuracy of the pitch estimate (Morise *et al.*, 2016). Hence we estimate the pitch from the Differentiated Electro Glotto Graph signal, based on which the spectral envelope, aperiodicity, are estimated using the WORLD vocoder. The pitch contour is denoted by $p[n]$, where n is the frame index. A sample pitch contour is shown in Fig. 1, where n_1 and n_2 denote the two vowel-consonant boundaries in a chosen VCV. Let the pitch drop p_d and the percentage of the pitch drop $p_{\%}$ in the C-region due to SFI is given by

$$p_d = p_b - p_{\min}, \quad p_{\%} = \frac{p_d \times 100}{p_b},$$

where $p_b = (p[n_1] + p[n_2])/2$ and $p_{\min} = \min_{n_1 \le m \le n_2} p[m]$. The goal is to modify the contour $p[n]$ by varying $p_{\%}$. We propose an affine transformation based modification for this purpose.

3.2 Modifying pitch contour

For the VCV perception study, we modify the pitch contour within the C-region. First we divide the pitch contour into two regions using the location of the $p_{\min}$ (n_c in Fig. 1). We modify pitch in these two regions using two different affine transforms such that the values of the pitch at the boundaries are preserved. Suppose we need to modify the pitch drop from $p_{\%}$ to $q_{\%}$. For this, the two parts of the C-region's pitch contour are separately modified using affine transformation as follows:

$$g^{q_{\%}}[n] = \begin{cases} s \times p[n] + (1-s) \times p[n_1] & \text{if } n_1 \le n \le n_c \\ t \times p[n] + (1-t) \times p[n_c] & \text{if } n_c \le n \le n_2, \end{cases} \tag{1}$$

where $s = (p[n_c] - p'_d - p[n_1])/(p[n_c] - p[n_1])$, $t = (p[n_c] - p'_d - p[n_2])/(p[n_c] - p[n_2])$ and $p'_d = p_b \times q_{\%}/100$. The modified contours with $q_{\%} = 2p_{\%}$, $q_{\%} = 24$, and $q_{\%} = 0$ are illustrated in Fig. 1. It is clear from the figure that the pitch values are continuous at the boundaries n_1, n_2 as well as at n_c. The values given by $g^{q_{\%}}[n]$ are used as the modified pitch contour for synthesizing the VCV stimuli.

4. Experiments and results

4.1 Experimental setup

Given the audio and EGG signals for a VCV recording, we extract the pitch contour $p[n]$ from the EGG signal using the DIO method (Morise *et al.*, 2010) for every 1 ms ($N_s = 16$). The modified pitch contour $g^{q_{\%}}[n]$ is obtained following the steps in Sec. 3.2.

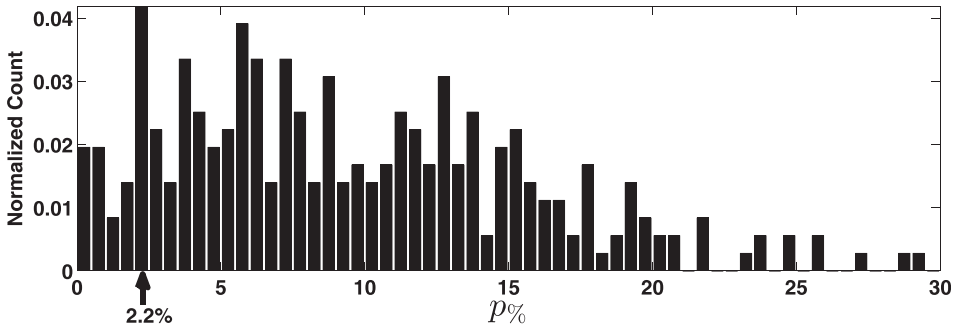

Fig. 2. Histogram of the $p_{\%}$ across 275 VCV samples used in this work.

The synthesized VCV sample is obtained using the modified contour $g^{q_{\%}}[n]$ and the original spectral envelope and aperiodicity from the WORLD vocoder. This is denoted by $x_q{}^{0}\!/_{0}$. The synthesized VCV using $p[n]$, the spectral envelope, and aperiodicity obtained from the WORLD vocoder is denoted by x_{AS}.

Different values of $q_{\%}$ for pitch modification is chosen based on the typical $p_{\%}$ in the dataset used in this work. The histogram of the $p_{\%}$ for all 275 VCV samples is shown in Fig. 2. It is clear from the figure that $q_{\%} = 2.2$ is the most frequently occurring value. In addition to $q_{\%} = 2.2$, we consider a large value of $q_{\%} = 24$ as only five VCVs have $p_{\%}$ above this. We also consider $q_{\%} = 0$ and $q_{\%} = 2p_{\%}$, i.e., the double of the $p_{\%}$ value of a given VCV sample. The corresponding synthesized speech is denoted by $x_{2.2}$, x_{24}, x_0, $x_{2p\%}$, respectively.

4.2 Listening test

We conducted subjective evaluations to assess the naturalness of the synthesised VCVs using a MUSHRA test (ITU, 2003), which allowed us to evaluate multiple samples in a single trial without breaking the task into many pairwise comparisons. The listening tests are conducted with a Direct Sound (Allentown, PA) EX-29 extreme isolation headphone. Prior to the listening test, each listener is asked to adjust the volume so that the sound played through the headphone is at his/her comfortable listening level. A MATLAB based graphical user interface is created where all five synthesized audio files (x_{AS}, $x_{2.2}$, x_{24}, x_0, $x_{2p_{\%}}$) are presented for each VCV and the original audio is used as the hidden reference. We consider a total of 30 listeners (15 male and 15 female) aged between 20 and 34 yrs for the MUSHRA test. None of the listeners were reported to have any hearing defects. Each listener was asked to listen to 30 VCVs, picked from 275 samples, such that each VCV is scored by at least three listeners. The listeners were asked to rate each stimulus from 0 (extremely bad) to 100 (perfect: same as the reference natural speech), and they were also instructed to give a rating of 100 to exactly one of the five stimuli in every set.

4.3 Results

The naturalness score for each VCV averaged over all listeners is shown in Fig. 3. The average scores are greater than 90 for all cases, which indicate that the naturalness of the synthesized VCVs is close to that of the original one. The natural speech has the highest mean score. As there is a difference in the average scores for synthesized VCV samples with different modified pitch contours, we perform the Analysis of Variance (ANOVA) (Gelman *et al.*, 2005) for differences among group means of x_{AS} and all other synthesized cases ($x_{2.2}$, x_{24}, x_0, $x_{2p_{\%}}$) for each VCV. This is done to measure the

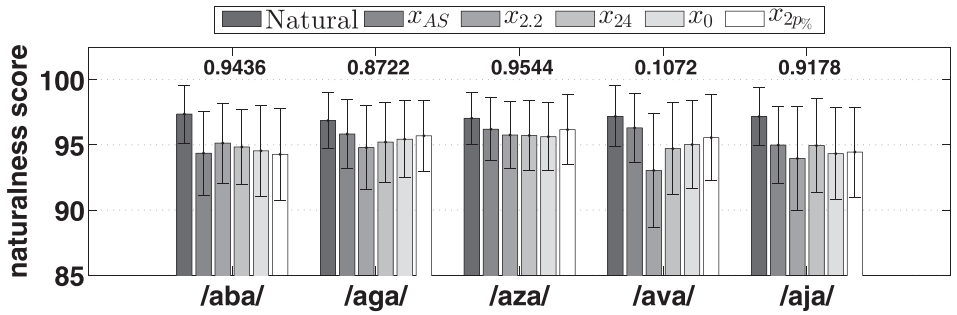

Fig. 3. Comparison of average scores given by the listeners and error bar indicate the 95% confidence interval. The *p*-value from the ANOVA analysis is shown at the top of the bar plots for each VCV.

distortion because of only the pitch modification. Even though there are differences in the mean scores of the synthesized stimuli, the *p-values* (>0.1) suggest that there is no statistically significant difference between the naturalness of the x_{AS} and that of different synthesized VCVs with pitch modification. This shows that the pitch change, because of SFI, is not critical for the naturalness of the VCV perception.

5. Conclusions

We study the effect of level-II SFI on the VCV naturalness perception for five different consonants in the context of the vowel /a/. We first decompose the VCV audio into pitch, spectral envelope, and aperiodicity. Following this, we modify the minimum pitch value in the C-region using an affine transformation based algorithm. The VCV audio is re-synthesized using the modified pitch contour and the original spectral envelope and aperiodicity. The listening test showed that no modification, as well as modification up to $2p_{\%}$, causes any statistically significant change in the perception of VCV naturalness. This shows that the pitch drop because of the SFI is not critical for the naturalness of a VCV. Our future works include similar studies on the role of SFI on the VCVs with different vowels, spontaneous speech and emotional speech perception.

References and links

Båvegård, M. (**2009**). "A study of voice source interaction," Working Papers Linguistics **43**, 38–41.

Chi, X., and Sonderegger, M. (**2007**). "Subglottal coupling and its influence on vowel formants," J. Acoust. Soc. Am. **122**(3), 1735–1745.

Ewan, W. G., and Ohala, J. J. (**1979**). "Can intrinsic vowel f0 be explained by source/tract coupling?," J. Acoust. Soc. Am. **66**(2), 358–362.

Fant, G. (**1971**). *Acoustic Theory of Speech Production: With Calculations Based on X-ray Studies of Russian Articulations* (Walter de Gruyter, Berlin, Germany), Vol. 2.

Fant, G., and Liljencrants, J. (**1979**). "Perception of vowels with truncated intraperiod decay envelopes," STL-QPSR **1**(1979), 79–84.

Gelman, A. (**2005**). "Analysis of variance—Why it is more important than ever," Annals Stat. **33**(1), 1–53.

ITU (**2003**). "1534-1: Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunication Union.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (**2008**). "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3933–3936.

Lucero, J. C., Lourenço, K. G., Hermant, N., Van Hirtum, A., and Pelorson, X. (**2012**). "Effect of source–tract acoustical coupling on the oscillation onset of the vocal folds," J. Acoust. Soc. Am. **132**(1), 403–411.

Mittal, V. K., Yegnanarayana, B., and Bhaskararao, P. (**2014**). "Study of the effects of vocal tract constriction on glottal vibration," J. Acoust. Soc. Am. **136**(4), 1932–1941.

Morise, M. (**2016**). "D4C, a band-aperiodicity estimator for high-quality speech synthesis," Speech Commun. **84**, 57–65.

Morise, M., Kawahara, H., and Nishiura, T. (**2010**). "Rapid f0 estimation for high-SNR speech based on fundamental component extraction," IEICE Trans. Inf. Syst. (Japanese edition) **J93-D**(2), 109–117.

Morise, M., Yokomori, F., and Ozawa, K. (**2016**). "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. Inf. Syst. **E99-D**(7), 1877–1884.

Nord, L., Ananthapadmanabha, T., and Fant, G. (**1984**). "Signal analysis and perceptual tests of vowel responses with an interactive source filter model," STL-QPSR **25**(2–3), 25–52.

Ohala, J. J., and Eukel, B. W. (**1987**). *Explaining the Intrinsic Pitch of Vowels* (Foris, Providence, RI), pp. 207–215.

Stevens, K. N. (**1971**). "Airflow and turbulence noise for fricative and stop consonants: Static considerations," J. Acoust. Soc. Am. **50**(4B), 1180–1192.

Stevens, K. N. (**2000**). *Acoustic Phonetics* (MIT Press, Cambridge, MA).

Taylor, P. (**2009**). *Text-to-Speech Synthesis* (Cambridge University Press, Cambridge, United Kingdom).

Titze, I., Riede, T., and Popolo, P. (**2008**). "Nonlinear source–filter coupling in phonation: Vocal exercises," J. Acoust. Soc. Am. **123**(4), 1902–1915.

Titze, I. R. (**2008**). "Nonlinear source–filter coupling in phonation: Theory," J. Acoust. Soc. Am. **123**(5), 2733–2749.

Titze, I. R., and Alipour, F. (**2006**). *The Myoelastic Aerodynamic Theory of Phonation* (National Center for Voice and Speech, Salt Lake City, UT).

Titze, I. R., and Palaparthi, A. (**2016**). "Sensitivity of source–filter interaction to specific vocal tract shapes," IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**(12), 2507–2515.

Vilkman, E., Laine, U. K., and Koljonen, J. (**1991**). "Supraglottal acoustics and vowel intrinsic fundamental frequency; An experimental study," Speech Commun. **10**(4), 325–334.

VoceVista. http://www.vocevista.com/electroglottograph/ (Last viewed April 1, 2018).

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (**1999**). "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*.