INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Frequency Patterns of Individual Speaker Characteristics at Higher and Lower Spectral Ranges

Zhao Zhang[1], *Ju Zhang*[2], *Ziyu Zhu*[1], *Yujie Chi*[1], *Kiyoshi Honda*[3], *Jianguo Wei*[1]

[1]College of Intelligence and Computing, Tianjin University, China
[2]Huiyan Technology (Tianjin) Co., Ltd., Tianjin, China
[3]The Academy of Tianjin University, Hefei, China

zhao_zhang@tju.edu.cn, jianguo@tju.edu.cn

Abstract

Acoustic characteristics of speech exhibit variability across individuals, while preserving shared phonetic information to listeners. In this paper, the general time-frequency pattern of individual speaker characteristics is discussed based on our previous research. The main target here is set at speaker-specific acoustic effects of the vocal tract in both higher and lower frequency ranges. To address the under-explored phenomena, two experiments were conducted. Firstly, simulations based on the transmission line model are used to explore how resonances in higher frequencies vary with different hypopharyngeal-cavity shapes. Secondly, speech signals emitted from the mouth and nostrils are recorded separately to observe potential factors for spectral irregularity in lower frequencies. From our findings, a time-frequency model of individual speaker characteristics is proposed that provides insights into how individuality is manifested in speech spectral patterns.

Index Terms: speaker characteristics, hypopharyngeal-cavity resonances, nasal-cavity resonances

1. Introduction

Speaker characteristics, being biophysical in nature, play an essential role in communication as they are transmitted together with literal spoken messages. Given the variability of speech signals in such utterances, it is challenging to determine how the vocal organs express a distinct code of individuality within dynamic acoustic streams of speech.

The most common acoustic features described in the past for speaker characteristics are the covariation of the fundamental frequency (F0) and formant frequencies of vowels, which can be perceived as characteristics of gender and age. Studies such as [1][2] have explored the diversity of F0 among males and females, and [3][4] have investigated how formant frequencies vary across speakers of different genders and ages. In addition to natural prosodic variations, phonatory variations have been found to cause gender differences in voice quality [5][6], while articulatory variations shown by [7][8] create richer speaker diversity. However, it is worth noting that these accounts do not fully cover the inter-speaker variation among the same gender and age. In addition to the above characteristics, a series of studies have found stable spectral evidence that is highly related to the vocal-tract variations among speakers.

In the higher-frequency region above 2.5 kHz, the hypopharyngeal cavities have been identified to create acoustic effects of speaker identity [9][10]. In male voices, the laryngeal cavity generates a stable spectral peak at about 3 kHz, while the zero-pole pair above 4 kHz is caused by the piriform fossa as side branches [11]. Our question is: male speakers exhibit a localized higher-frequency pattern, whereas the acoustic influence of the hypopharynx in female voices appears diffuse, which is unsolved even by a few attempts [12][13]. In the lower frequency region, the relevant organs to modify vowel spectra are the subglottal and nasal tracts. The resonance of the subglottal tract creates spectral dips at frequencies corresponding to three subglottal formants [14]. The resonance of the nasal cavity modifies vowel spectra by means of transvelar nasal coupling in non-nasal sounds [15][16], which also manifests the lower frequency region in a speaker-specific way. The next question is: the nostril radiation signals reflect nasal-cavity resonance, but how the signals modulate oral output signals of non-nasal sounds remains unclear.

In this paper, we aim at summarizing the static individual characteristics found in spectral patterns of vowels in conjunction with our prior research. Two experiments were conducted to elucidate the two aforementioned issues. The first experiment utilizes simulations based on a transmission line model to explore the effects of varying hypopharyngeal-cavity shapes on resonances in higher frequencies. The second experiment examines signals emitted from the mouth and nostrils separately to identify possible factors influencing speaker-specific spectral irregularity in lower frequencies. Lastly, we introduce a time-frequency table that provides a comprehensive overview of the static individual speaker characteristics in both higher and lower frequencies

2. Issues on the Characteristics in Higher and Lower Frequencies

2.1. On the Higher Frequencies

The hypopharynx consists of the laryngeal cavity and bilateral piriform fossa as shown in Fig. 1, forming the lowest portion of the vocal tract. Those small cavities exhibit relatively stable shapes during speech production [11], exhibiting consistent higher-frequency spectral patterns of regional resonance.

In the hypopharynx, the morphology of the laryngeal cavity resembles that of a Helmholtz resonator, comprising a small cavity (ventricles) and a lengthy neck (vestibular tube). [17] has indicated a gender variation in the cavity geometry, being longer in male than in female. Earlier work confirmed the male larynx cavity as an emitter of sharp resonance at about 3 kHz [18].

The other structure, the piriform fossa is situated above the esophageal entrance at the posterolateral aspects of the larynx, and acts as a side branch of the vocal tract. For male speakers, it generates a zero-pole pair above 4 kHz [13].

Thus, the hypopharynx plays a crucial role in shaping the higher frequency spectra of vowels, with the peak of laryngeal cavity resonance and the zero-pole pattern of the piriform fossa.

10.21437/Interspeech.2023-1803

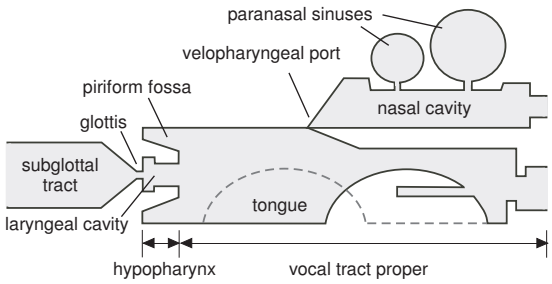

Figure 1: *Schema of the vocal tract and other airways. Modified from [19].*

While the acoustic effect of the male hypopharynx can be effectively modeled, the impact in female cases is more complex. From [12], both positive and negative peaks are distributed over a wider and higher frequency range with significant variability across subjects and vowels in female data.

2.2. On the Lower Frequencies

The lower frequency region contains a diverse spectral components. The subglottal tract, as depicted in Fig. 1, is responsible for generating spectral dips in the lower frequency range with its second resonance at 1.3–1.5 kHz [14]. Given that the subglottal tract is relatively stable across individuals, its resonances may serve as an indicator of speaker characteristics. The nasal cavity is another fixed conduit, and it impacts non-nasal sounds through the transvelar nasal coupling. It is reported that the resonance of the nasal cavity in isolation displays a double-peak pattern below 1 kHz [20]. This resonance pattern was found to vary across speakers in our study [21]. However, the specific influence of nasal-cavity resonance on oral vowel sounds remains unclear. This is partly because natural recorded vowels are the outcome of acoustic interference between the oral and nasal channels.

3. Experimental setup

3.1. Acoustic simulation with varying hypopharyngeal-cavity shapes

To explore the underlying mechanisms responsible for the observed variations in female spectral patterns in the higher frequency regions, acoustic simulations were conducted utilizing vocal-tract data obtained from a female speaker producing vowel /a/ after [22]. The image processing involved the extraction of the vocal tract and perioral surface from MRI data, followed by the calculation of the vocal-tract area function using the procedure detailed in [23].

In this experiment, the impact of variations in the shapes of the laryngeal cavity and single piriform fossa on spectra was examined separately. Each of their area functions was approximated as a single tube with a uniform cross-sectional area, referred to as "larynx tube" and "piriform duct" hereafter. The simplified vocal-tract schematic and its corresponding area function for the female data are depicted in Fig. 2.

The acoustic simulation was conducted based on the transmission line model [24] with varying lengths and areas of the larynx tube and piriform duct. For comparison, the frequency response of the vocal tract proper (without the hypopharyngeal cavities) was measured first. The following settings of experiments were employed to analyze the corresponding frequency response patterns.

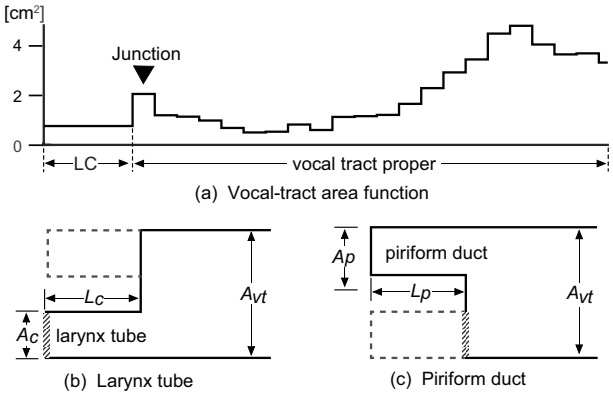

Figure 2: *Simplified vocal-tract configuration (female case). (a) Vocal-tract area function with the larynx tube and the junction to calculate the area ratio. (b) Simplified model for larynx tube. (c) Simplified model for piriform duct. The hatched boundary in (b) and (c) indicates the location of excitation source.*

Length change: The effect of changing lengths of the larynx tube and piriform duct (L_p, L_c) in the model was examined with the scale factors 1.5, 1.25, 1.0, 0.75, and 0.5, while the larynx/ piriform area (A_c/A_p) was kept constant. The initial L_p and L_c are both 2 cm.

Area change: The impact of altering cross-sectional areas for the larynx tube and piriform duct (A_p, A_c) was investigated by manipulating the area ratio of the larynx tube and piriform duct to their junction to the hypopharynx (A_c/A_{vt}, A_p/A_{vt}), as depicted in Fig. 2 (a). The initial A_p,A_c and A_{vt} are $1.27\,\mathrm{cm}^2$, $0.31\,\mathrm{cm}^2$ and $2.22\,\mathrm{cm}^2$.

3.2. Separated recordings from the mouth and nostrils

To investigate potential factors contributing to spectral irregularity in lower frequencies, speech signals were recorded separately from the mouth and nostrils for spectral analysis. To achieve this, a solid panel is positioned between the upper lip and nostrils to isolate the two signals. Two omnidirectional microphones (AT9904, Audio-Technica) were positioned on the upper and lower sides of the panel for simultaneous recording.

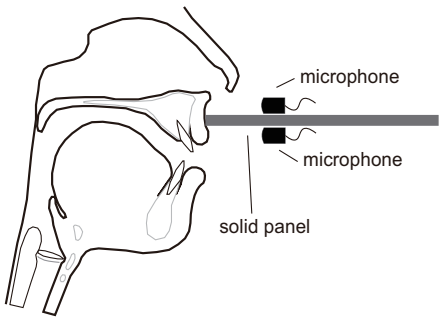

Figure 3: *Recording set up for separated oral and nostril output signals.*

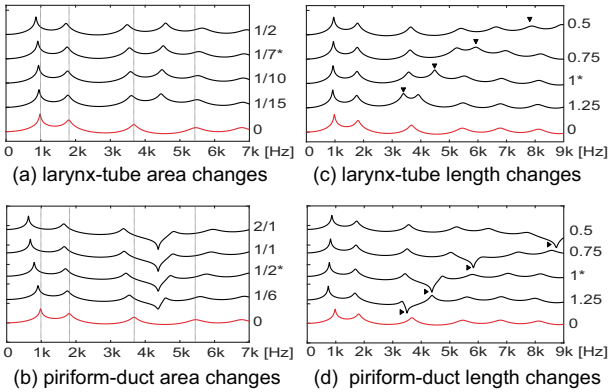

Figure 4: *Results of acoustic simulations. (a) and (c) show the frequency responses with varied areas and lengths of the larynx tube, respectively. (b) and (d) show the frequency responses with varied areas and lengths of the piriform duct, respectively. The area ratios and length scale factors used in the simulations are defined in section 3, and the corresponding values are shown on the right side of the figures. The area ratios marked with an asterisk (*) represent the original configuration of the speaker. In each panel, the bottom curve is the transfer function of the vocal tract proper.*

Normal speech signals were also recorded using the same microphone in a separate session. All the data were sampled at 44100 Hz and stored on an audio recorder (PCM-D10, SONY).

This experiment is a part of our ongoing study [21], examining nasal-cavity contribution to non-nasal sounds during six Mandarin vowels ($/a/, /o/, /e/, /i/, /u/, /y/$) and a short Chinese sentence to derive the long-term average spectrum (LTAS). Given the frequent occurrence of small velopharyngeal openings during the production of low vowels, the sentence was constructed without vowel /a/ and consisted of the phonetic sequence ($/wo \quad i\varepsilon \quad wu \quad yi \quad wei \quad yu \quad yi \quad wu \quad yi \quad yi/$).

To compare the oral, nostril, and natural signals, let the spectral envelope of the oral output signal be denoted by S_o, and that of the normal signal by S_n. By subtracting S_o from S_n ($S_n - S_o$), a subtraction envelope was calculated to identify the peak-dip distribution pattern in the spectra.

4. Results

4.1. Simulation results by tube/duct modifications

4.1.1. Result on the larynx tube

In Fig. 4(a), the resonance peak of the larynx tube in the female model is found at approximately 4.5 kHz. This peak induces shifts of the original vocal-tract resonance peaks (formants), causing lower formants to shift to lower frequencies and higher formants to shift to higher frequencies. Changes of the larynx-tube resonance (4.5 kHz) are minute because of the lack of the ventricle in the model. As the larynx tube's cross-sectional area increases, the influence of the larynx tube becomes more evident over the entire frequency range.

The effect of larynx tube length changes is demonstrated in Fig. 4(c), where the tube resonance peak shifts with the length. The longer larynx tube produces a resonance peak at lower frequencies, and the peak moves to higher frequencies and becomes flatter and broader as the length decreases. In the extreme case where the length of the original larynx tube (2 cm) is halved, the resonance peak widens and shifts to a very high frequency, resembling the resonance of the vocal tract with an additional length.

4.1.2. Result on the piriform duct

The results in Fig. 4(b) and 4(d) exhibit a clear zero-pole pattern in all simulations when the piriform duct is added. In Fig. 4(b), alteration of the piriform duct's cross-sectional area from narrow to wide leads to deepening and widening of the anti-resonance, which in turn influences the entire spectrum by shifting other resonances (formants) to both sides of the spectrum, centered around the anti-resonance at 4.3 kHz. In extreme scenarios where the piriform duct's cross-sectional area exceeds the junction area of the vocal tract proper, notable shifts of the first three formants could be observed. Figure 4(d) indicates that the location of the anti-resonance is determined by the length of the piriform duct, with the longer piriform duct driving the anti-resonance at lower frequencies, and vice versa. It is also found that the shorter piriform duct exhibits the broader range of variability in the higher frequency region. In the case in Fig. 4(d) for the piriform duct shortened from the scale factor 1.25 (2.5 cm) to 1 (2 cm), the anti-resonance frequency shifts from 3501 Hz to 4371 Hz, indicating a variation of 870 Hz. As for shortening the piriform duct from the scale factor 1 (2 cm) to 0.75 (1.5 cm), a more substantial shift is observed in the anti-resonance frequency, from 4371 Hz to 5841 Hz, with a variation of 1470 Hz. The findings indicate that the shorter duct length results in the more pronounced shift of anti-resonance frequency.

4.2. Comparison of oral, nostril, and natural signals

The oral and nostril signals recorded in the experiment represent resonances of the vocal tract (with no nasal cavity) and the nasal cavity (with a closed velopharyngeal port), respectively. Figure 4(a) and 4(b) show spectra envelopes from the separated channels in vowels /i/ and /u/, respectively. The oral signal spectra resemble the natural ones with minor changes in vowel formants, while the nostril signal spectra show low-frequency energy concentration, exhibiting strong dumping with a sharp spectral decay from the low-frequency peak (around 750 Hz) to higher frequencies.

In LTAS for the separated channels depicted in Fig. 4(c), a distinct pole-zero-pole pattern is also observed below 1 kHz, reflecting an antiresonance of a large sinus cavity.

A comparative analysis is made on vowel data employing spectral subtraction of normal and oral signals. As summarized in Fig. 4(d), the subtracted spectra demonstrate a uniform distribution of peaks and dips below 2 kHz, albeit with slightly varying intervals corresponding to different vowel phonemes.

5. Discussions

This study dealt with the spectral characteristics signaling speaker identity in the higher and lower frequency ranges mainly by focusing on unsolved questions regarding hypopharyngeal resonance and nasal cavity resonance.

In the simulation experiment on the hypopharynx, the acoustic effects of varying shapes of the larynx tube and piriform duct were explored separately in a preliminary manner. The results indicate that the resonances of both structures exhibit similar variation patterns that are specific to their respective shapes. The location of the resonance/anti-resonance in the spectra is mainly determined by the length of the larynx/piriform regions, with the longer structures creating lower-

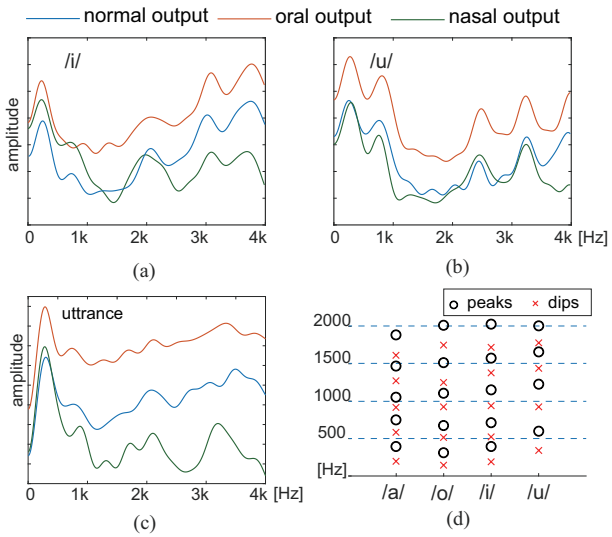

Figure 5: *Spectral results from oral and nostril recordings. (a), (b) Spectral envelopes from vowel /i/ and /u/. (c) LTAS envelopes from the sentence. (d) Spectra peak-dip distribution of subtracted spectral envelopes between normal and oral output signals of four vowels.*

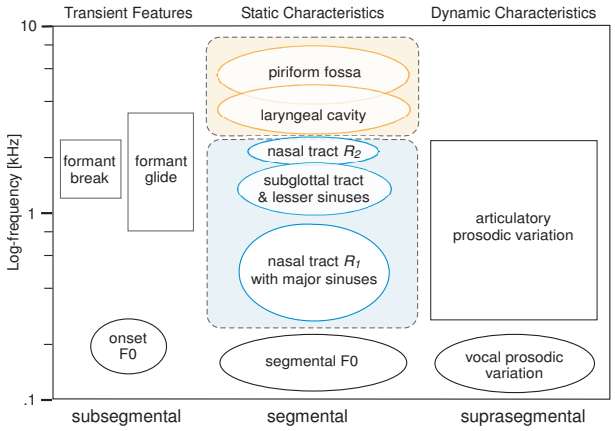

Figure 6: *Time-frequency chart of individual speaker characteristics. Topics of the study are color-surrounded for the higher and lower frequency components comprising static speaker characteristics in vowels.*

frequency resonances, and vice versa. The impact range of the resonance and anti-resonance is mainly influenced by the area configuration of the larynx/piriform region, with the larger areas producing deeper and wider resonances. When the hypopharynx resonances occur, the formant frequencies shift towards opposite sides of the spectrum, centered around each of the original hypopharynx resonances. The simulation results provide a potential explanation for the complex pattern observed in the higher frequency range of female speakers. Specifically, the shorter hypopharynx in females leads to the broader distribution of hypopharynx resonances in the higher frequency range. And the relative volume of the hypopharynx to the vocal tract proper is slightly larger in females, possibly resulting in a stronger interaction between the two structures. As a result, compared to male speakers, female speakers have more diverse individual variations, which leads to high-frequency differences among female speakers.

The spectral analysis of recordings from the mouth and nostrils identifies two peaks at approximately 240 Hz and 750 Hz. These peaks were observed due to a peak-splitting anti-resonance, a phenomenon previously reported in [20][25], and an irregular undulation above 1 kHz, consistent with [26]. The analysis of the peak-dip distribution of the subtracted spectra envelope suggested an assumption that the acoustical coupling of the nasal and oral tracts may lead to constructive and destructive interference effects, deriving a subtle modulation of peaks and dips as observed at low-level anti-formant regions in the 1–2 kHz frequency range.

Drawing upon our accumulated discoveries, we have compiled a time-frequency chart to summarize our findings as shown in Fig. 5 in sub-segmental, segmental, and suprasegmental features. Sub-segmental features refer to transient attributes, while suprasegmental features pertain to prosodic phrasal components. Our focus in this study is on the segmental features comprising static individual characteristics during vowel production. In the higher frequency range, males exhibit localized spectral patterns with a peak at about 3 kHz and a zero-pole pair above 4 kHz. In contrast, females demonstrate a greater variation distributed over a broader frequency range: Both laryngeal cavity resonance and anti-resonances of the piriform fossa can be observed across a wide frequency range from 4 kHz to as high as 8 kHz. For lower frequencies, individual speaker characteristics can be observed in (1) double-peaked first resonance of the nasal cavity with a large sinus (below 1 kHz), (2) spectral irregularity caused by the subglottal tract and smaller sinus cavities (1 – 2 kHz), and (3) the second nasal-cavity resonance (around 2 kHz). In the future, further research is needed to explore a more definitive frequency range for female high-frequency features based on finer morphological data.

6. Conclusion

This paper attempted at proposing a general time-frequency model for individual speaker characteristics. Two experiments were conducted to find possible explanations for two issues raised in our previous studies. Simulations based on a transmission line model were conducted, and the result suggested that the shorter and wider hypopharynx in female derives hypopharynx resonances in a wide range in higher frequencies with overlapped spectral components. The signals from the mouth and nostrils were recorded separately to be compared, with the result that nasal-cavity resonance influences the oral vowel sounds by adding a peak-dip-peak pattern below 1 kHz. Finally, we introduced a time-frequency chart that provides a comprehensive overview of the static individual speaker characteristics in both higher and lower frequencies. As details are further explored, it will be possible to provide technical assistance in speaker identification, spoof detection, and other applications.

7. Acknowledgements

This work was supported by National Key R&D Program of China(No.2020YFC2004103), Qinghai science and technology program (No. 2022-ZJ-T05), and the project of Tianjin science and technology program (No.21JCZXJC00190).

8. References

[1] H. Hollien, P. A. Hollien, and G. de Jong, "Effects of three parameters on speaking fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2984–2992, 1997.

[2] C. G. Henton, "Fact and fiction in the description of female and male pitch." *Language & Communication*, vol. 9, no. 4, pp. 299–311, 1989.

[3] P. A. Busby and G. L. Plant, "Formant frequency values of vowels produced by preadolescent boys and girls," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2603–2606, 1995.

[4] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[5] C. G. Henton and R. A. Bladon, "Breathiness in normal female speech: Inefficiency versus desirability." *Language & Communication*, vol. 5, no. 3, pp. 221–227, 1985.

[6] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466–481, 1997.

[7] K. N. Stevens, "Sources of inter-and intra-speaker variability in the acoustic properties of speech sounds," in *Proceedings of the 7th international congress of phonetic sciences*, 1971, pp. 1596–1607.

[8] L. L. Koenig, "Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 5, pp. 1211–1228, 2000.

[9] S. Furui, "Perception of voice individuality and physical correlates," *Trans. Tech. Com. Psycho. Physio. Acoust.*, 1985.

[10] L. Zhang, K. Honda, J. Wei, and S. Adachi, "Regional Resonance of the Lower Vocal Tract and its Contribution to Speaker Characteristics," in *Proc. Interspeech 2020*, 2020, pp. 1391–1395.

[11] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustical Science and Technology*, vol. 26, no. 1, pp. 16–26, 2005.

[12] J. Li, K. Honda, J. Zhang, and J. Wei, "Individual difference and acoustic effect of female laryngeal cavities," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[13] C. Zhang, K. Honda, J. Zhang, and J. Wei, "Contributions of the piriform fossa of female speakers to vowel spectra," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[14] S. M. Lulich, J. R. Morton, H. Arsikere, M. S. Sommers, G. K. Leung, and A. Alwan, "Subglottal resonances of adult male and female native speakers of american english," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2592–2602, 2012.

[15] H. Suzuki, T. Nakai, J. Dang, and C. Lu, "Speech production model involving subglottal structure and oral-nasal coupling through closed velum." in *ICSLP*, vol. 90, 1990, pp. 437–440.

[16] J. Dang, J. Wei, K. Honda, and T. Nakai, "A study on transvelar coupling for non-nasalized sounds," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 441–454, 2016.

[17] J. Zhang, K. Honda, J. Wei, and T. Kitamura, "Morphological characteristics of male and female hypopharynx: A magnetic resonance imaging-based study," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 734–748, 2019.

[18] J. Sundberg, "Articulatory interpretation of the "singing formant"," *The Journal of the Acoustical Society of America*, vol. 55, no. 4, pp. 838–844, 1974.

[19] K. Honda, "Physiological processes of speech production," *Springer Handbook of Speech Processing*, pp. 7–26, 2008.

[20] M. Havel, S. Becker, M. Schuster, T. Johnson, A. Maier, and J. Sundberg, "Effects of functional endoscopic sinus surgery on the acoustics of the sinonasal tract," *Rhinology*, vol. 55, no. 1, pp. 81–89, 2017.

[21] Z. Zhu, Y. Chi, Z. Zhang, H. Kiyoshi, and J. Wei, "Transvelar nasal coupling contributing to speaker characteristics in non-nasal vowels," in *Proc. Interspeech 2023 (this conference)*, 2023.

[22] K. Honda, T. Kitamura, H. Takemoto, S. Adachi, P. Mokhtari, S. Takano, Y. Nota, H. Hirata, I. Fujimoto, Y. Shimada *et al.*, "Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 13, no. 4, pp. 443–453, 2010.

[23] H. Takemoto, K. Honda, S. Masaki, Y. Shimada, and I. Fujimoto, "Measurement of temporal changes in vocal tract area function from 3d cine-mri data," *The Journal of the Acoustical Society of America*, vol. 119, no. 2, pp. 1037–1049, 2006.

[24] X. Zhou, Z. Zhang, and C. Espy-Wilson, "Vtar: a matlab-based computer program for vocal tract acoustic modeling," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2543–2543, 2004.

[25] S. Maeda, "The role of the sinus cavities in the production of nasal vowels," in *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7. IEEE, 1982, pp. 911–914.

[26] S. Takeuchi, H. Kasuya, and K. Kido, "A study on the effects of nasal and paranasal cavities on the spectra of nasal sounds," *The Journal of the Acoustical Society of Japan*, vol. 33, no. 4, pp. 163–172, 1977 (in Japanese).