INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Transvelar Nasal Coupling Contributing to Speaker Characteristics in Non-nasal Vowels

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

[1]College of Intelligence and Computing, Tianjin University, China
[2]The Academy of Tianjin University Hefei, China

2020244147@tju.edu.cn, yujiechi@qq.com, williezz@163.com, khonda@sannet.ne.jp, jianguo@tju.edu.cn

Abstract

Nasal-cavity structure is stable in speech and varied across speakers, which potentially gives rise to speaker characteristics. Many studies have reported the acoustic contribution of the nasal cavity for nasal and nasalized sounds with velopharyngeal port opening. However, nasal-cavity resonance does emerge in non-nasal vowels through transvelar nasal coupling, which results in non-negligible modifications to non-nasal vowel spectra. In this study, nasal and oral output sounds were separately recorded during non-nasal utterances, and spectral analysis was conducted. The results indicate clear inter-speaker variability in two spectral measures below 2 kHz: frequency location of double-peaked first nasal-cavity resonance and inconsistent distribution of minor dips above the first resonance. It was also observed that nostril outputs modulate oral output signals to lower the first formant frequency of naturally produced non-low vowels, which also exhibited varied degrees across speakers.

Index Terms: nasal-cavity resonance, transvelar coupling, speaker characteristics

1. Introduction

Speech signals are characterized by non-stationary spectra, as evidenced by changes in vowel formants on a spectrogram. Such signals also transmit stationary information arising from individual differences of the speech organs. This notion is of particular importance in understanding speaker identity in speech sounds, since resonance in the vocal tract is known to be unique to each individual.

Earlier studies have revealed that individual characteristics are found in the higher frequency regions above 2.5 kHz [1]. The higher-frequency features are known to be caused by the hypopharynx consisting of three small cavities (laryngeal cavity and bilateral piriform fossae) [2]. In male speakers, the laryngeal cavity causes a resonance peak at around 3 kHz, and the piriform fossae introduce zero-pole pairs above 4 kHz [3, 4, 5].

In contrast, the lower frequency features are underexplored except for formant frequency spacing that reflects vocal-tract length [6]. This is because the frequency region below 2.5 kHz is occupied by moving vowel formants during speech, which smears fixed cavity resonances. The possible organs introducing stationary low frequency features are the subglottal tract and nasal cavity. The subglottal system acts as a vocal-tract branch when the glottis is open, and its resonance introduces dips and peaks on vowel spectra near its resonant frequencies [7, 8]. The nasal cavity also affects low-frequency spectra of vowels. In the production of nasal and nasalized vowels, the nasal cavity is jointed to the vocal tract via velopharyngeal port opening (VPO) [9]. VPO introduces a zero or two on low frequency spectra, which can be tuned by altering the port size (e,g., the split first formant of vowel /a/) [10]. Further, additional zero-pole pairs are known to derive from the paranasal sinuses and asymmetry of the nasal passages [11, 12, 13, 14]. The nasal cavity is not independent from the vocal tract even for non-nasal sounds such as oral vowels and voiced stops. This is because sound pressure variation in the vocal tract behind oral constriction is transmitted into the nasal cavity through the velum even when the velopharyngeal port is closed. This acoustic effect is known as transvelar nasal coupling [15] or transpalatal nasalance [16]. The greater effect of transvelar coupling have been observed in higher vowels [17]. However, individual variations of nasal-cavity resonance and its nostril radiation in non-nasal vowels have rarely been explored.

In this paper, we focus on the spectral characteristics of nasal-cavity resonance in non-nasal vowels by recording oral and nostril outputs separately. To do so, a new experimental setup was designed to suppress oral or nostril output signal at recording the target signal. In addition, subglottal resonance signals were monitored for their potential spectral overlap. Then, spectral analysis was conducted to reveal the nasal resonance details and their effects on vowel spectra.

2. Experimental setup

In the experiment, the nostril or oral output sounds was recorded separately using acoustic devices to suppress non-target signals. To record weak nostril output in oral vowel production, an oral silencer and nose-mask microphone were constructed as shown in Fig. 1. To record oral output signals without nostril output, a nose mask was used as a nasal silencer.

2.1. Oral and nasal silencers and nose-mask microphone

To record nostril output sounds, the oral silencer unit is constructed to minimize the interference by oral output sound. This unit consists of a mouthpiece, reflectionless tube, and sound absorber box. At recording, subjects kept the plastic mouthpiece with clay contacted on the perioral tissue. The oral output sound is led to the sound absorbing box via a reflectionless tube (100-cm long) having a corn-shaped absorber at the end. This output end of the tube is inserted to a wooden enclosure, which is internally lined with sound insulator (rubber sheet) and absorber (urethane sponge) for the minimal sound transmission.

For clean recording of weak nostril output, a nose-mask microphone unit is used during utterances. This unit is crafted by modifying an air-filtering nose mask (NF-B01, Maixingren). The filter inside was replaced by cotton gauze, and vent holes were relocated to avoid contamination by oral output. The elastic fringe of the mask keeps hermetic contact on the whole nose.

To record oral output sound, the nasal silencer is used, which is essentially similar to the nose-mask microphone unit,

10.21437/Interspeech.2023-637

except for having no hole for a microphone. The mask was filled with cotton gauze, with vent holes on the top of the mask.

By measuring effects of the silencer units, oral shielding by the oral silencer was 21 dB at 1 kHz, and the regain of nasal radiation loss by the nasal silencer was 16 dB at 1 kHz.

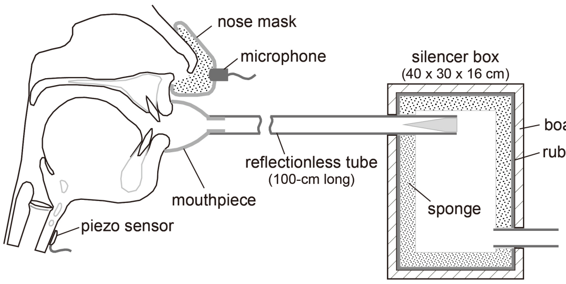

Figure 1: *Components of the experimental setup to record nostril output sounds. A mouthpiece, reflectionless tube, and sound absorber box form an oral silencer unit. Nostril output is recorded using a nose-mask microphone unit.*

2.2. Recording devices

Two omni-directional microphones (AT9904, Audio-Technica) were used for recording acoustic signals. One unit is attached to an ear-mount bracket for recording normal speech. The other was used for oral and nostril output recording separately. A piezoelectric vibration sensor was employed to record subglottal tract resonance by placing it below the cricoid cartilage. All those signals were connected to a sound console (AG03, YAMAHA) to be stored on a PC.

3. Recording and spectral analysis

3.1. Data recording

This study examines the acoustic characteristics of six Chinese vowels (/a/, /o/, /ɤ/, /i/, /u/, /y/) and a Chinese short sentence. In Chinese, the vowel /a/ is a low vowel. /o/ and /ɤ/ are half-high vowels. The last three are high vowels. Those oral vowels were analyzed for spectral envelopes. The short sentence was used to calculate the long-term average spectrum (LTAS). Since small velopharyngeal opening is common for low vowels [18], the short sentence was made excluding vowel /a/, forming '我也误以为语义无意义' (/wo iɛ **wu** yi wei yu yi **wu yi yi**/; 'I also mistakenly think the meaning is meaningless' in English). To obtain smooth spectral envelopes from LTAS, bold words were emphasized to distribute harmonics over frequencies.

Four types of audio data were recorded with the setups explained in section 2. The specific procedures and purposes are as follows:

(1) Natural sound: Recorded with the ear-mount microphone placed between the mouth and nose.

(2) Oral output: Recorded by the ear-mount microphone, while wearing the nasal silencer (Fig. 2(a)).

(3) Free nostril output (with no mask): Recorded by the ear-mount microphone placed near the nostrils together with the oral silencer (Fig. 2(b)).

(4) Nose-mask sound: Recorded by a nose-mask microphone with the oral silencer (Fig. 2(c)). This is to record non-radiated nostril output with minimal contaminations.

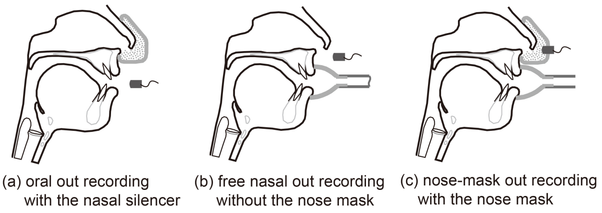

Figure 2: *Experiment setups corresponding to oral outputs and two types of nostril outputs. Microphones were placed to keep a 2-cm distance from each output end.*

3.2. Subjects for the experiment

Subjects for the experiment were three females (ZY, YJ, QW) and three males (MS, ZZ, JB), aged between 23 and 36. All of them have no surgery on vocal organs and speak the standard Mandarin. They repeated each vowel and sentence three times. Recordings were conducted in a soundproof room with the temperature kept at 23°C.

3.3. Preprocessing with a high-pass filter

All the signals were sampled at 16000 Hz with 16-bit resolution. The subglottal signals and nostril output signals from the nose-mask microphone showed a sharp spectral decay toward higher frequencies. An FIR high-pass filter, with a cut-off frequency (*Fc*) of 2 kHz and 6 dB attenuation at *Fc*, was applied for spectral flattening before the pre-emphasis (0.93). The stopband attenuation is 27 dB.

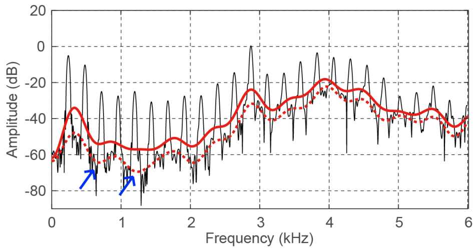

Figure 3: *Spectrum of natural vowel /i/ from a female speaker (ZY). The solid line is the envelope calculated by a conventional cepstral analysis, and the dashed line from the inverted Imai's method.*

3.4. Spectral envelope estimation

Cepstral analyses were conducted to obtain spectral envelopes by averaging steady sections of vowels using half-overlapping 100-msec Hann windowing. To confirm the analysis being free from harmonic attractions, the obtained spectra were compared by those from the inverted Imai's method mentioned in [19] as modified from [20]. This custom method is to suppress harmonics and trace the baselines of FFT spectra that correspond to tract resonance excited by glottal airflow noise.

Figure 3 compares spectral curves obtained from the two cepstral methods; conventional and inverted Imai's. The two dips at the low frequency are sometimes more obvious on the dotted line (inverted Imai's), which are close to the subglottal resonances of the subject ZY at 584 Hz and 1432 Hz.

3.5. Long-term average spectrum

Long-term average spectrum (LTAS) was applied to the short sentence containing non-low vowels. This method is to obtain stationary spectral details by suppressing the effects of harmonics and vowel formants. The average spectra were calculated using the same half-overlapping window as used for vowel analysis. Also, the subglottal formants are obtained on LTAS.

4. Results

4.1. Amplitude variation across vowels

The normal sounds (with no silencers) and oral output signals (with nasal silencer) for the six vowels exhibited a natural pattern of intrinsic vowel intensity: The high vowels (/i/, /u/ and /y/) were weaker than the other vowels, and the low vowel /a/ was the largest.

The nostril output signals (with nose-mask microphone and oral silencer) demonstrated the highest amplitude in vowel /i/ among the non-low vowels. In vowel /a/, two different amplitude patterns were observed across speakers, that is the larger vs. smaller amplitude in comparison to that in vowel /i/. The first type of augmented amplitude in /a/ is seen in Fig. 4(a), while the second type of smaller amplitude is also found as shown in Fig. 4(b).

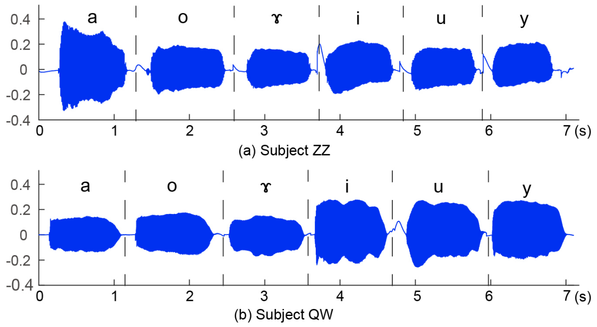

Figure 4: *Waveforms of nose-mask outputs across six vowels showing two different amplitude patterns between /i/ and /a/ in two speakers.*

4.2. Nostril output LTAS from nose-mask microphone

The LTAS curves of nostril output sounds recorded with and without the nose-mask microphone are compared in Fig. 5. The signal amplitude from the nose-mask microphone is markedly greater, with the peaks and dips below 3 kHz at the same frequencies, which supports the effectiveness of the nose-mask microphone for the low frequency range.

4.3. Nasal-cavity resonance and subglottal formant

The LTAS envelopes on nostril output signals are assumed to represent spectral characteristics of nasal-cavity resonance caused by transvelar nasal coupling. Figure 6 summarizes the main peaks and zeros in the nose-mask output spectra for six subjects. The two peaks with a zero in between (*P1-D1-P2*) are consistently observed for all subjects below 1 kHz. The first peak is gathered at 300 Hz. The second peak varies among different speakers. Above the *P1-D1-P2* region, a less obvious second dip (*D2*) appears below 2 kHz at varying frequencies across speakers except for the subject ZZ. Small dips (shown by smaller crosses) appear between 1–2 kHz for several subjects. The third peak (*P3*) is consistent for all speakers with distribution around 2.5 kHz. The 2nd subglottal formant (*SgF2*) obtained from the short utterance distributes at around 1.4 kHz. In QW (female), the peaks, zeros, and *SgF2* tend to be higher than those in other subjects. The covariation of *SgF2* with *P1-D1-P2* across subjects is not obvious.

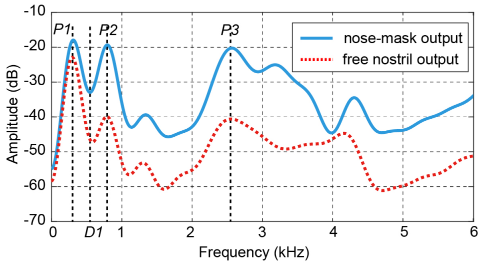

Figure 5: *Comparison of spectral envelopes on LTAS of nose-mask output (with nose-mask microphone) and free nostril output (without nose-mask unit) from subject YJ.*

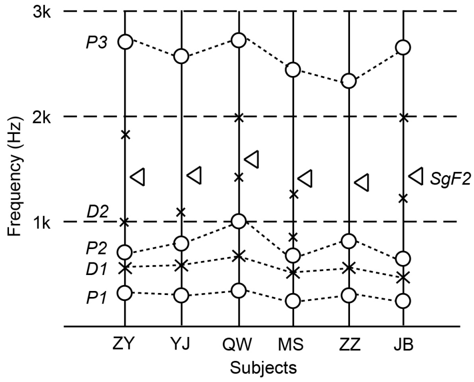

Figure 6: *Peaks (Pn) and dips (Dn) on LTAS from nose-mask output sounds for six subjects. SgF2 is the 2nd formant of subglottal signals extracted from normal outputs.*

4.4. Nasal-cavity resonance in non-nasal vowels

Figure 7 shows observed states of nasal-cavity resonance for all six subjects. Each panel includes spectral envelopes for two vowels (/a/ and /i/) and LTAS on the short utterance. For these speakers, the main peaks and dips of the nostril output sounds appear in vowel /i/, which also resemble the peaks and dips in LTAS. The spectra for vowel /a/ show a qualitative difference. The lowest peak and zero in /a/ sometimes agree with those in /i/ and LTAS, while the overall spectral patterns disagree. This difference in /a/ is assumed due to a different excitation mechanism of nasal-cavity resonance in /a/: The velopharyngeal port in low vowels tends to open to various degrees across speakers. For example, in data from the subject ZZ shown in Fig. 7, the nostril output in /a/ shows significantly different peaks from other vowels. The nasal peaks in /i/ and LTAS are at 300 Hz, 820 Hz, and 2344 Hz, while the peak values for /a/ are at 761 Hz, 1684 Hz, and 2848 Hz.

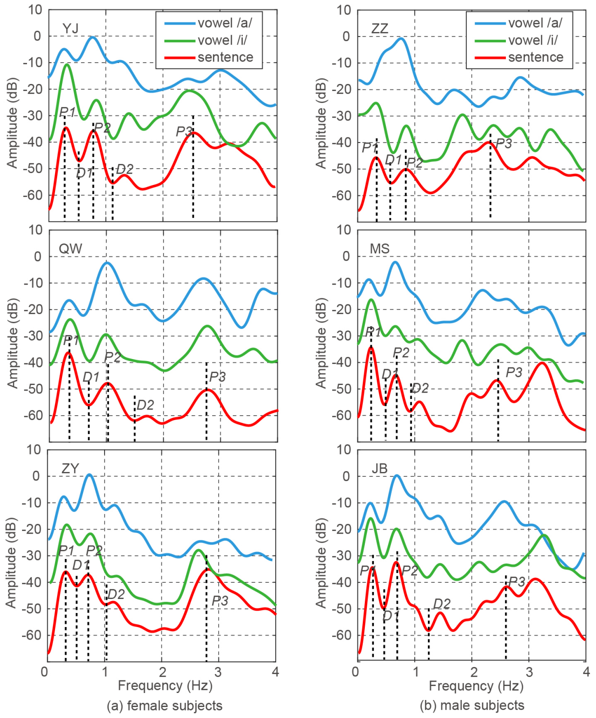

Figure 7: *Nostril-out spectra in vowels and LTAS on the short sentence obtained by the nose-mask microphone. In each panel, the line is blue for /a/, green for /i/, and red for LTAS.*

4.5. Formant frequency differences between natural and oral-output vowels

The effect of the nostril output on natural vowel spectra was examined by comparing the first two formants (F1 and F2) in the high (/i/ and /u/), mid-high (/o/), and low (/a/) vowels. As shown in Fig. 8, F1 frequency of most of natural vowels was lower than that in the oral output sounds to various degrees. This tendency was evident especially for high vowels. For vowel /i/, the average fall of F1 frequency for female speakers is 12.6%, and that for male speakers is 1.9%. For vowel /u/, the average decrease is 9.0% for females and 4.8% for males. No obvious differences are found in F2 frequency of the vowels examined.

5. Discussions

The spectral characteristics of nasal-cavity resonance in non-nasal speech were investigated as low-frequency components of static speaker characteristics. By developing unique devices, the nostril and oral output sounds were recorded separately using the oral silencer and nasal silencer, respectively. Below, we discuss the spectral characteristics of nostril output sounds below 3 kHz and their effect on oral vowel spectra.

The nasal-cavity resonance in LTAS of nostril output sounds shows consistent double peaks with a dip in between at around 500 Hz in the region below 1 kHz (Fig. 6). According to [11, 21], this double-peak pattern is due to a zero introduced on the first nasal resonance, and the zero is assumed as the anti-resonance caused by the maxillary sinuses. The dip and peak at 1-2 kHz are presumably introduced by the paranasal sinuses with smaller volumes [14, 22]. Another eminent peak distributed at 2.5 kHz is probably the second nasal resonance, which is also observed in [17] by measuring intranasal sound pressure.

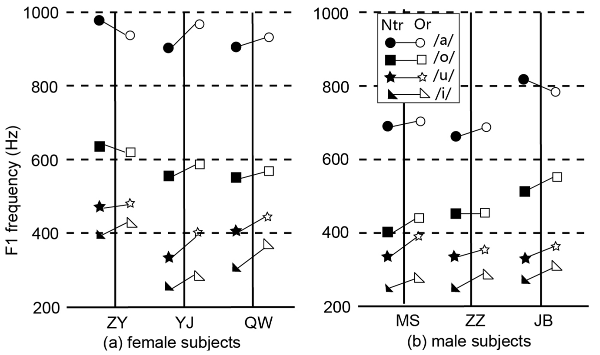

Figure 8: *The 1st formant frequencies (F1) of normal and oral-out vowels plotted for (a) female and (b) male speakers. Filled and empty marks are 'natural' (Ntr) and 'oral out' (Or) vowels, respectively.*

The nasal resonances were notably observed on LTAS of the nostril output of non-low vowels. Among those, the larger amplitude is revealed in vowel /i/, which is due to the greater transvelar nasal coupling. The situation differs in vowel /a/: the amplitude is sometimes larger than in vowel /i/, which is inconsistent with [15]. This deviation is likely to be caused by velopharyngeal port opening in vowel /a/, resulting in eminent nasal-cavity resonance in this vowel. The port opening in low vowels is a common finding [18], which supports our observation.

The transvelar coupling is one of the main mechanisms modulating non-nasal sounds, especially the vowels with the closed port [17]. The sounds transmitted into the nasal cavity radiates to the open space through the nostrils, causing interference with the sound radiated from the m outh. In this process, lower frequencies are augmented and the first formant is lowered as seen in normal vowels, which is consistent with the simulation results in [23]. In this study, the greater impact is seen for female speakers. All those processes are thought to contribute to generating lower frequency speaker characteristics.

6. Conclusion

Oral vowel spectra are assumed to form a smooth curve with formants and anti-formants, whereas it is common to observe complex peak-dip modulation in real data. One of the reasons is the fact that the velum transmits sound energy into the nasal cavity. Due to this transvelar nasal coupling, the low-frequency resonance in the nasal cavity augments lower frequencies in high vowels near F1 with a frequency shift. Thus, non-nasal vowels reflect nasal resonance that gives rise from excitation near the velum, either via velum vibration or narrow port opening. The mechanism of how the nostril output blends with the oral output in non-nasal speech should be investigated further.

7. Acknowledgement

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

8. References

[1] S. Furui, "Perception of voice individuality and physical correlates," *Tech. Rep. Hear. Acoust. Soc. Jpn.*, 1985.

[2] 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.

[3] S. Fujita and K. Honda, "An experimental study of acoustic characteristics of hypopharyngeal cavities using vocal tract solid models," *Acoustical Science and Technology*, vol. 26, no. 4, pp. 353–357, 2005.

[4] 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.

[5] 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, pp. 1391–1395.

[6] W. T. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," *The Journal of the Acoustical Society of America*, vol. 102, no. 2, pp. 1213–1222, 1997.

[7] X. Chi and M. 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.

[8] 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.

[9] K. N. Stevens, *Acoustic Phonetics*. MIT press, 1998.

[10] P. Birch, B. Gümoes, S. Prytz, A. Karle, H. Stavad, and J. Sundberg, "Effects of a velopharyngeal opening on the sound transfer characteristics of the vowel [a]," *Speech, Music and Hearing*, pp. 9–15, 2002.

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

[12] S. Maeda, "Acoustics of vowel nasalization and articulatory shifts in French nasal vowels," *Nasals, Nasalization, and the Velum*, vol. 5, pp. 147–167, 1993.

[13] J. Lindqvist-Gauffin and J. Sundberg, "Acoustic properties of the nasal tract," *Phonetica*, vol. 33, no. 3, pp. 161–168, 1976.

[14] T. Pruthi, C. Y. Espy-Wilson, and B. H. Story, "Simulation and analysis of nasalized vowels based on magnetic resonance imaging data," *The Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3858–3873, 2007.

[15] J. Dang and K. Honda, "Investigation of the acoustic characteristics of the velum for vowels," in *Proc. ICSLP*, vol. 94, 1994, pp. 603–606.

[16] E. L. Bundy and D. J. Zajac, "Estimation of transpalatal nasalance during production of voiced stop consonants by noncleft speakers using an oral-nasal mask," *The Cleft Palate Craniofacial Journal*, vol. 43, no. 6, pp. 691–701, 2006.

[17] 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.

[18] K. L. Moll, "Velopharyngeal closure on vowels," *Journal of Speech and Hearing Research*, vol. 5, no. 1, pp. 30–37, 1962.

[19] Z. Zhang, K. Honda, and J. Wei, "Retrieving vocal-tract resonance and anti-resonance from high-pitched vowels using a rahmonic subtraction technique," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7359–7363.

[20] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electronics and Communication*, vol. 62, no. 4, pp. 10–17, 1979, in Japanese.

[21] 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.

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

[23] J. Dang and K. Honda, "An improved vocal tract model of vowel production implementing piriform resonance and transvelar nasal coupling," in *Proceeding of Fourth International Conference on Spoken Language Processing. (ICSLP)*, vol. 2, 1996, pp. 965–968.