Physics of Body-Conducted Silent Speech – Production, Propagation and Representation of Non-Audible Murmur

Makoto Otani[1], *Tatsuya Hirahara*[2]

[1] Faculty of Engineering, Shinshu University, Japan
[2] Faculty of Engineering, Toyama Prefectural University, Japan
otani@cs.shinshu-u.ac.jp, hirahara@pu-toyama.ac.jp

Abstract

The physical nature of weak body-conducted vocal-tract resonance signals called non-audible murmur (NAM) were investigated using numerical simulation and acoustic analysis of the NAM signals. Computational fluid dynamics simulation reveals that a weak vortex flow occurs in the supraglottal region when uttering NAM; a source of NAM is a turbulent noise source produced due to a vortex flow. Furthermore, computational acoustics simulation reveals that NAM signals attenuate 50 dB at 1 kHz consisting of 30-dB full-range attenuation due to air-to-body transmission loss and –10-dB/octave spectral decay due to a sound propagation loss within the body, which roughly equals to the measurement results.

Index Terms: non-audible murmur, glottal flow, turbulent noise source, body-conducted sound, attenuation characteristics

1. Introduction

A non-audible murmur (NAM) which is a very weak speech sound produced without vocal cord vibration has been researched to in the development of a silent speech communication tool [1, 2]. The NAM can be detected using a specially designed microphone, called a NAM microphone [3, 4]. This is a small condenser microphone covered with a soft impression material such as soft silicon or urethane elastomer, which provides better impedance matching between biological soft tissues and the condenser microphone diaphragm. With the NAM microphone attached to the neck surface close behind an ear, a NAM can be detected as a body conducted voice. The NAM, which is inaudible even for people nearby, can be audible when detected by the NAM microphone. This enables the development of human-to-human and human-to-machine interfaces whose inputs are inaudible voices, thereby providing a "silent" communication tool. The NAM microphone will also be able to revive the speech communication of those with vocal cord problems caused by laryngeal cancer, nerve disorders or muscle diseases.

While development of the NAM microphone has improved the detection of body-conducted speech sounds with a stethoscope, the physical nature of body-conducted speech sounds has been left underexplored. In this work, production and propagation mechanisms of the NAM are investigated numerically using computational fluid dynamics (CFD) simulation and computational acoustic/vibration simulation, respectively. CFD simulation is performed using the finite volume method (FVM) and three-dimensional (3D) vocal tract shape models. Acoustic/vibration simulation is performed using the finite difference time domain (FDTD) method and two-dimensional (2D) head model. Models of vocal tract and head shape are obtained by magnetic resonance imaging (MRI) scans. Furthermore, numerical results of acoustic attenuation and propagation characteristics are validated with results obtained by acoustic analysis of the NAM signals sensed using the NAM microphone.

2. Production Mechanism of NAM

2.1. MRI Scan during NAM Production

Vocal tract shapes were obtained for NAM production by using MRI scans to construct 3D vocal tract shape models. Three adult male subjects between 22 and 30 years old participated in the MRI scan. The subjects uttered NAMs of the vowel /i/, which was chosen because it is the easiest one for subjects to stabilize an articulatory position. In frontal vowels, the tongue is held in contact with the palate, thereby increasing the articulation repeatability [5]. The vocal tract shape during the voice production was scanned using phonation-synchronized MRI scans [6]. SHIMADZU-Marconi MAGNEX ECLIPS 1.5T at ATR-BAIC was used. The scanning sequence was RF-FAST. The echo time (TE) was 3 ms, and the repetition time (TR) was 12 ms. For subjects 1 and 2, MR images were recorded on a saggital plane; each image was 512 x 512 pixels, field of view (FOV) was 256 x 256 mm, slice thickness was 1 mm, and spatial resolution was 0.5 x 0.5 x 1.0 mm. For subject 3, MR images were recorded on a coronal plane; the image size was 512 x 512 pixels, FOV was 128 x 128 mm, slice thickness was 1 mm, and spatial resolution was 0.25 x 0.25 x 1.0 mm.

2.2. 3D Vocal Tract Shape Model

We constructed 3D vocal tract shape model consisting of triangular meshes from the MRI images of one subject who resulted in the "clearest" images. The procedure was as follows. First, the MR images were edge-reinforced. Next, they were binarized based on the edge-reinforced sagittal, coronal, and horizontal images. Finally, 3D vocal tract shape models were constructed by interpolating from the binarized volume data.

Compared to the original MR images, the 3D vocal tract shape model had 0.42-mm error in the maximum height and 0.43-mm error in the maximum width of piriform fossae. These errors are smaller than the MR spatial resolution, which is 0.5 mm, indicating that 3D model reasonably reproduces vocal tract shape.

2.3. Numerical Simulation of Glottal Flow during NAM Production

A 3D incompressible simulation was performed using a commercial 3D CFD solver (FLUENT 6.3.26, ANSYS,

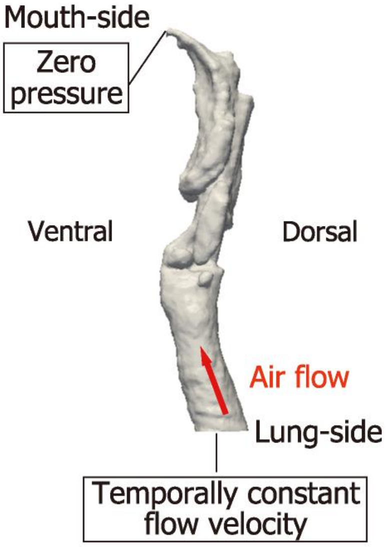


Figure 1: *Sagittal view of vocal tract shape model. Air flows from lung-side to mouth-side. Temporally constant flow velocity given at lung-side boundary. Zero pressure given at mouth-side boundary.*

Inc. [7]) based on the finite volume method. The governing equations are the Navier-Stokes equations. The turbulence model used was large-eddy simulation (LES, Smagorinski Lilly, $C_s = 0.1$). The time difference scheme was the implicit second-order method (time step: 50 [μs]). The spatial difference scheme used was bounded central differencing. Tetrahedron cells were automatically generated from the vocal tract shape model by a mesh generator (ICEM-CFD, ANSYS, Inc.). The boundary condition at the lung-side opening of the vocal tract was given as a temporally constant flow velocity, whose value was determined so that it matched the values of the volume flow rate described below. The boundary condition at the mouth-side opening of the vocal tract exit was given as pressure P (= 0). The air density and viscosity was 1.225 [kg/m^3] and 1.79 x 10^{-5} [Pa·s], respectively. Figure 1 illustrates the numerical condition.

An expiratory volume flow rate for the NAM production was determined according to previous research [8, 9]. Netsell *et al.* reported that the expiratory volume flow rate of ordinary voice production is between 100 and 200 [ml/s] [8]. Rubin *et al.* reported that the expiratory volume flow rates of weak and strong whispered voices are double and triple that of the ordinary voice, respectively [9]. Therefore, the volume flow rate of a weak whispered voice was estimated to be approximately 300 [ml/s]. Because the NAM has a much weaker energy than a weak whispered voice, the volume flow rate for NAM was determined to be 100 [ml/s] in this work. The number of cells and nodes of the vocal tract shape model are approximately 365,000 and 125,000, respectively.

2.4. Results

Powell suggested that vorticity $\boldsymbol{\omega}$ [1/s] (= rot($\mathbf{u}$), where $\mathbf{u}$ [m/s] is the volume flow velocity vector) relates to a vortex sound source; namely, a sound wave arises where a vortex exists [10]. Figure 2 shows instantaneous vorticity magnitude $|\boldsymbol{\omega}|$ distributions on sagittal and coronal sections at 25 ms after the start time of simulation. The colors represent the values of the vorticity magnitude ranging from 0 to 1,000, as the color bar indicates. The result shows that the vorticity magnitudes are larger in the supraglottal space, indicating that the vortex flows yield turbulence noise sources in NAM production as in a whispered voice production. Note that turbulent noise sources in NAM production do not always occurs at supraglottal region near the glottis; namely, similar to a whisper production, sources for some types of consonant such as fricative or plosive, *e.g.* /s/ or /k/, are produced due to vortex flows arising at an alveolar or a velar. Clarifying how the produced vortex flow relates to the frequency characteristics of the NAM signals remains for a future work.

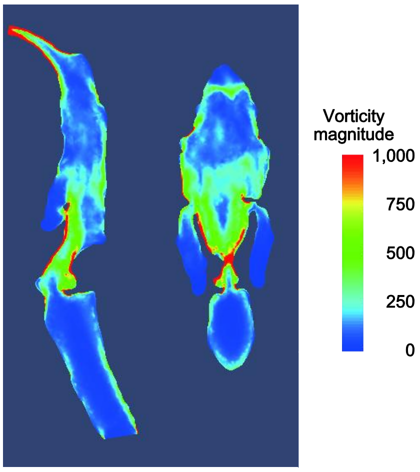


Figure 2: *Instantaneous vorticity magnitude (in [1/s]) distributions on sagittal and coronal sections for NAM. Color bar ranges from 0 to 1,000.*

3. Propagation Mechanism of NAM

Sound originating in the vocal tract propagates from the air column inside the tract to the body tissue and can be detected at the body surface. The shortest propagation distance from the vocal tract to the neck surface, where a NAM microphone is commonly placed, is approximately 70 mm. To clarify the NAM transfer characteristics, we numerically investigated sound propagation from the vocal tract to the neck surface using the 2D FDTD method and a head model constructed on the basis of MRI scans.

3.1. Numerical Simulation of NAM propagation

3D geometrical data of a human head during production of the vowel sound /e/ were obtained using phonation-synchronized MRI scans [6]. An image on the midsaggital plane was extracted to produce a 2D head model. For simplicity, a homogeneous head model was generated, *i.e.* the head was approximated as being composed of only soft tissue. Finally, the vocal tract was replaced by a simplified model with a rectangular duct with a cross-section 30 mm wide. This approximation of the vocal-tract shape was made because a scanned vocal tract is extremely narrow in the vicinity of the larynx, and it would require a grid geometry that would be too fine to be simulated by the FDTD method and the current computational resources. Cyber Logic Wave 2000 Pro was used as the FDTD solver. Details of the simulation are described in [11].

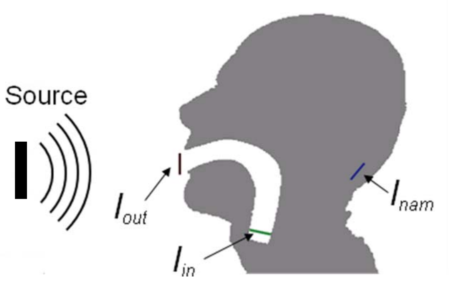


Figure 3: *Geometry of simulated region including head, sound source, and receivers.*

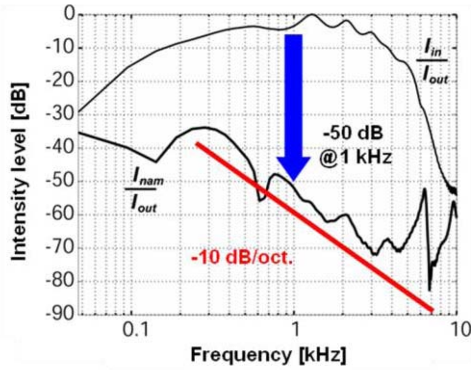


Figure 4: *Vocal-tract transfer functions I_{in}/I_{out} and I_{nam}/I_{out} corresponds to sound intensity level observed at NAM microphone assuming sound radiating from mouth has flat frequency characteristics.*

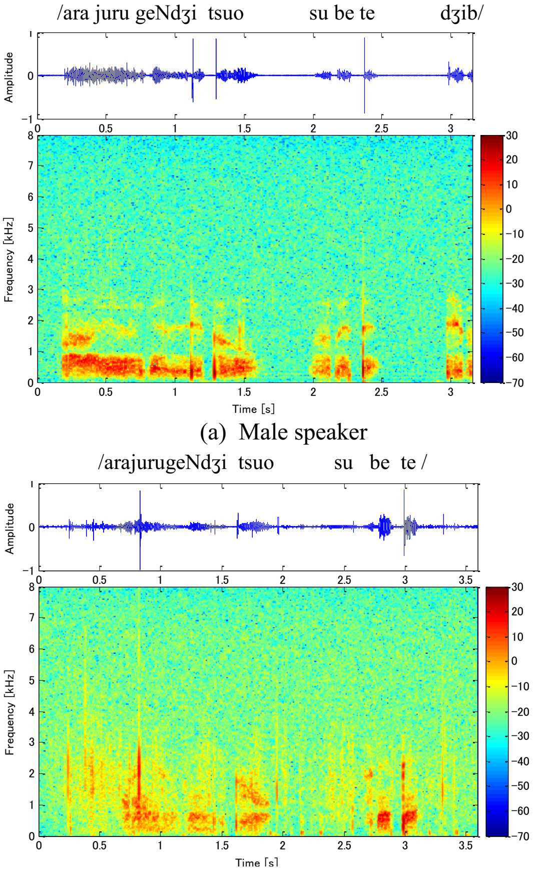
3.2. Results

Figure 3 shows the geometry of the simulated region including the head, the sound source, and the receivers. The sound source was located in front of the mouth, assuming a reciprocal theorem, as was a pulse-driven vibrating plate. Receivers were located at the inlet and outlet of the vocal tract and on the neck behind the ear, where a NAM microphone would normally be attached. These receivers are referred to as "*in*", "*out*" and "*nam*", respectively. The sound intensity levels at each receiver I_{in}, I_{out}, and I_{nam}, were calculated, and then I_{in}/I_{out} and I_{nam}/I_{out} were calculated as shown in Fig. 4 The I_{in}/I_{out} corresponds to the vocal-tract transfer functions, showing spectral peaks, *i.e.* formants, at 0.5, 1.4, 2.1, and 3.0 kHz. The $I_{nam}/I_{out} = (I_{nam}/I_{in})/(I_{out}/I_{in})$ corresponds to the sound intensity level observed at *nam* assuming that the sound radiating from the mouth has flat frequency characteristics. The transmission loss of sound as it passed from the air in the vocal tract into the soft tissue and the propagation loss through the soft tissue $I_{nam}/I_{in} = (I_{nam}/I_{out})/(I_{in}/I_{out})$ were –50 dB at 1 kHz. The loss consists of 30-dB full-range attenuation due to air-to-soft-tissues transmission loss and –10 dB/octave spectral decay due to propagation loss in the soft tissues.

4. Acoustic Characteristics of NAM signal

4.1. Recording Procedure

The NAM signals were recorded in a soundproof room. Six male and seven female adult participants read 50 ATR phoneme-balanced Japanese sentences [12] in NAM mode of weak whispering. A soft-silicone type NAM microphone was used. It was attached to the neck behind the ear below the mastoid process and fixed in place with a neckband. The microphone was carefully positioned so as to optimize detection of the NAM signals. The signals were recorded using a solid state recorder (PMD670; Marantz) at a sampling rate of 48 kHz with 16-bit resolution. During the recording, the NAM phonation mode and the signal level were monitored by an audio expert to maintain the quality of the recorded signals. When intense noise was detected, the speaker was instructed to reread the sentence. In total, 80-min NAM signals for 640 sentences were recorded. A very weak whispering sound, which would radiate from the mouth during the NAM production, does not interfere with the NAM signal recorded at the neck surface.



(a) Male speaker

(b) Female speaker

Figure 5: *Waveform and spectrogram of NAM signals for (a) male and (b) female speaker recorded with soft-silicone type NAM microphone.*

4.2. Waveforms and Spectrograms

The waveforms and spectrograms of two example recorded signals are shown in Fig. 5. A male and female speaker each read a Japanese sentence, "arajuru geNdʒitsuo [subete]"; its meaning is "all the reality". Despite the low-amplitude nature of the NAM signals, speech segments are identifiable in the waveform. The background noise level, however, was high, and a number of impulsive pop noises overlapped the NAM signals, suggesting that the signal-to-noise ratio (SNR) of the signals was low. The pop noises were generated either by contact between the microphone and skin or by contact of the articulatory organs when producing plosives and fricatives. Some of this type noise can be reduced by level thresholding as the amplitude of the noise was generally larger than that of the signals. After such pop noises are reduced, the speech

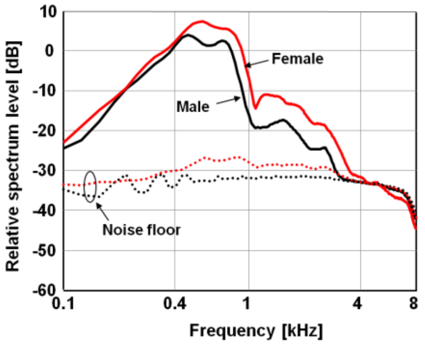


Figure 6: *Long-term average spectra (LTAS) of NAM signals resampled at 16 kHz.*

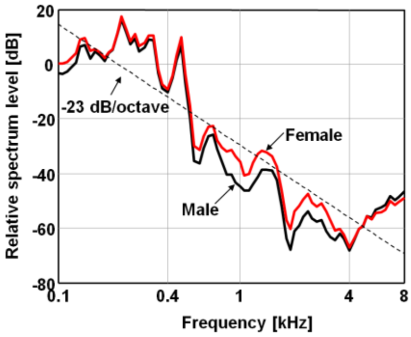


Figure 7: *Long-term average spectra (LTAS) of underlying NAM signals by subtracting frequency responses of NAM microphone and amplifier from LTAS shown in Fig. 6.*

segments of NAM signals can be automatically detected using their power and delta-cepstrum coefficients.

4.3. Signal to Noise Ratio

The SNRs of the NAM signals were significantly lower than those of ordinary speech signals obtained with a regular microphone. The mean SNR was 15 dB for NAM signals recorded at a sampling rate of 48 kHz. Reducing the sampling rate improved the SNR – it reached a maximum of 21 dB at a sampling rate of 2 kHz.

4.4. Long-Term Average Spectra

The long-term average spectra (LTAS) of NAM signals resampled at 16 kHz are shown in Fig. 6: the black and red solid lines indicate the mean LTAS for six male and seven female speakers, respectively, and the dotted line of each color shows the noise floor of the corresponding signal. The LTAS data show that the bandwidth of the signals was 3 kHz.

The frequency responses of the NAM microphone and amplifier played a role in the overall frequency characteristics of the NAM signals. Subtracting those frequency responses from the LTAS lead to the LTAS of underlying NAM signal, as shown in Fig. 7. The LTAS show that the spectrum rolls off at –23 dB/octave above 300 Hz. According to the source-filter theory of speech production, the LTAS of ordinary voice has an –6-dB/octave roll-off. As the radiation characteristics from the lips are +6 dB/octave, the spectral tilt of the source was estimated to be –12 dB/octave for ordinary voice. Thus, the underlying NAM signal involves a –11 dB/octave spectral decay due to voice sound transfer through the body, which roughly equals to –10 dB/octave suggested by the numerical simulation.

5. Discussion

The NAM microphones can detect a silent body-conducted voice as mentioned. Using NAM microphones with appropriate speech conversion or speech recognition system enables developments of a "silent" speech interface from human to human/machine. Additionally, not limited to detecting "silent" voice, NAM microphones are capable of detecting various body-conducted biological sounds originating in a human body. Such applications of NAM microphones would contribute to realizing biological-sound monitoring for the purpose of medical use or life log.

6. Conclusion

The physical nature of weak body-conducted vocal-tract resonance signals called non-audible murmur (NAM) was investigated. The CFD simulation reveals that a weak vortex flow occurs in supraglottal region when uttering NAM; similarly to an ordinary whispering, a source of NAM is a turbulent noise source produced due to a vortex flow arising in a supraglottis region. The acoustics simulation reveals that NAM signals attenuate 50 dB at 1 kHz; this attenuation consists of 30-dB full-range attenuation due to air-to-body transmission loss and –10 dB/octave spectral decay due to a sound propagation loss within the body, which are roughly equal to the spectral decay observed in measured characteristics of NAM signal.

7. Acknowledgements

Part of this work is supported by SCOPE of the Ministry of Internal Affairs and Communications of Japan. The authors would appreciate Dr. Y. Nakajima and Mr. S. Shimizu for their supports in NAM signal recording and data analysis, respectively.

8. References

[1] Nakajima, Y. *et al.*, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," *Proc. IEEE ICASSP*, 708-711, 2003.

[2] Hirahara, T. *et al.*, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *Speech Com.* **52**(4): 301-313, 2010.

[3] Nakajima, Y. *et al.*, "Non-audible murmur (NAM) recognition," *IEICE Trans. on Information and Systems* **E89-D**(1): 1-8, 2006.

[4] Shimizu, S. *et al.*, "Frequency characteristics of several types of NAM microphones," *Acoust. Sci. & Tech.* **30**(2): 139-142, 2009.

[5] Kitamura, T. *et al.*, "Difference in vocal tract shape between upright and supine postures observations by an open-type MR scanner," *Tech. Rep. IEICE Speech* **104**(149): 1-6, 2004.

[6] Nota, Y. *et al.*, "A bone-conduction system for auditory stimulation in MRI," *Acoust. Sci. & Tech.* **28**(1): 33-38, 2007.

[7] ANSYS FLUENT Flow Modeling Software http://www.ansys.com/products/fluid-dynamics/fluent/

[8] Netsell, R. *et al.*, "Vocal tract aerodynamics during syllable productions: Normative data and theoretical implications," *J. Voice* **5**(1): 1-9, 1991.

[9] Rubin, A. *et al.*, "Laryngeal hyperfunction during whispering: reality or myth?" *J. Voice* **20**(1): 121-127, 2006.

[10] Powell, A., "Theory of vortex sound," *J. Acoust. Soc. Am.* **36**(19): 177-195, 1964.

[11] Otani, M. *et al.*, "Numerical simulation of transfer and attenuation characteristics of soft-tissue conducted sound originating from vocal tract," *Appl. Acoust.* **70**: 469-472, 2009.

[12] Abe, M. *et al.*, "Speech database user's manual," ATR Tech. Rep. TR-I-0166, ATR Interpreting Telephony Res. Lab., 1990.