ICIC Express Letters
Part B: Applications
Volume **14**, Number **9**, September **2023**

ICIC International ©2023 ISSN 2185-2766
pp. **951–959**

ANALYSIS FOR SKIN PROPAGATION SPEECH WITH THREE DIMENSIONAL DISPLACEMENT MEASUREMENT

Masashi Nakayama[1,*], Takumi Muto[2], Daiki Kawamoto[1] and Shunsuke Ishimitsu[2]

[1]Graduate School of Information Sciences
[2]Faculty of Information Sciences
Hiroshima City University
3-4-1, Ozuka-higashi, Asaminami, Hiroshima 731-3194, Japan
ishimitu@hiroshima-cu.ac.jp
[*]Corresponding author: masashi@hiroshima-cu.ac.jp

Received January 2023; accepted April 2023

Abstract. *In the measurement of skin vibration on a face when a speaker utters or sings as speech, it is general to use a laser Doppler meter as a measurement method for one-dimensional displacement. However, considering the facial shape of a human body and the complexity of sound propagation on the body, it is desirable to use a measurement method with multidimensional and high-frequency resolution. Therefore, we propose to examine vibrations which are the propagation of the facial skin and conduction of the human body during utterance and singing, and then clarify the relationship with utterance by using three-dimensional displacement measurement with two axes as x and y. In this study, we confirmed frequency characteristics and signal intensities from skin propagation speech which is focused on the upper lip during an utterance, and clarified the relationship between utterance and skin vibration on the face and body. In addition, we also studied to extract the fundamental frequency from skin vibration during vocalization and investigate the relationship between speech and skin propagation speech focusing on the fundamental frequency.*

Keywords: Speech, Skin propagation speech, Three dimensional displacement, Formant frequency, Fundamental frequency

1. **Introduction.** Speech is one of the important methods for our daily conversation, and this sound is conducted from the speaker's mouth to the listener's ear via airborne. At the same time, the sound is also conducted by the speaker's body, such as by skin and bone, via a solid propagation, which is often called the body or Bone-Conducted Sound (BCS). Many researchers made investigations to propose BCS as a speech interface because it has noise-robust characteristics compared to airborne noise. On the other hand, there are few studies on the related outcomes of skin conduction sound because it is difficult to find out and show the benefit of clarifying its conduction mechanism.

By elucidating skin propagation on a human face, it is expected that singing training will find a mechanism for producing higher pitches and longer vocalizations. In addition, a relationship between vocalization manner and singing/spoken formants can be systematically clarified by speech, BCS, and skin vibration. Shortly, it will appear that an implantable interface for BCS is realized by a microchip implanted in the human body with electricity self-generation, and it is foreseeable that a technology for restoring sound from vibrations on there. Therefore, the authors predict the era of monitoring biological signals and performing voice communication that transcends various environments such as underwater without air, outer space, and high noise.

DOI: 10.24507/icicelb.14.09.951

However, few researchers investigated measurement of skin vibration sound using a laser vibrometer as one of the measurement methods from non-contact signal measurement. Tabatabai et al. experimented to measure signals that are muscle tension, tremor, heart sound, and voice with a single point of a vibrometer [1]. Measurement using a laser vibrometer or accelerometer was investigated to clarify the skin and body propagation of a human body and skin during speech vocalization and singing. Kitamura measured with a vibrometer when three Japanese singers vocalized Japanese vowel /a/ with a fixed fundamental frequency [2]. Avargel and Cohen proposed speech enhancement using vibrometer information and frequency characteristics improvement with the spectral gain modification and information [3]. Lee et al. proposed electronic skin for quantitative vocal recognition [4]. In this study, we proposed the idea of measuring the skin propagation during vocalization by sound vibration measurement target using a three-dimensional displacement measurement that can capture three-dimensional displacements on a human's face because the measurement can measure surface vibration using a multidimensional recording with high sampling frequency. It is expected to help elucidate signal propagation because human speech, body-conduction speech, and skin vibration sound are measured simultaneously. The interrelationship of the experimental outcomes for these conduction channels is discussed.

This manuscript consists of the following sections. Section 2 describes the difference in the skin vibration sound when measured by a laser Doppler meter as a conventional solution and in three-dimensional displacement measurement as the solution proposed here. Section 3 shows a comparison of frequency characterizations between signals, and Section 4 shows the dependence of vibration strength on axis directions. Section 5 shows measurement and improvement for the fundamental frequency with the estimation method. Finally, Section 6 makes conclusions and indicates future work.

2. Measurement of Skin Vibration as Skin Propagation Speech.

To measure skin vibration sound on a human's face, a scanning measurement with a laser Doppler meter is generally used as a conventional solution. Because this measurement uses the reflection of a laser on the skin, it is a one-dimensional displacement measurement in the vertical direction concerning this device. However, considering the complexity of human skin changes, a multidimensional measurement is required. Therefore, using a three-dimensional displacement measurement, we examine the way to measure the skin surface vibration and propagation. In three-dimensional displacement measurement, it is possible to image the surface using the high-speed capability of the imaging camera installed at different angles and to obtain three-dimensional displacement measurement at an arbitrary point within the imaging range. Three-dimensional displacement measurement has been used in measurements for testing and examination of advanced materials and parts because there is high speech scanning with precision at the time domain and spatial resolution [5]; however, a vibration measurement for human skin has not been attempted. Thus, the method is expected to enable three-dimensional displacement measurement for humans.

3. Experiment 1: Comparisons of Frequency Characteristics.

In general, during vocalization, in addition to airborne sound radiated from the lips and nostrils, there is also a body-conducted sound propagating through solids, such as skin, via the conduction mechanism. In this study, an experiment is conducted to find the differences in frequency characteristics of the speech, body-conducted speech, and skin vibration sound using three-dimensional displacement measurements.

3.1. Experimental setup.

Speech measurements were performed with a microphone, and the body-conducted speech was measured with an accelerometer and skin propagation with three-dimensional displacement measurement. Figure 1 shows a scene of sound recording. This experiment was performed indoors, in a part of the office dedicated to the

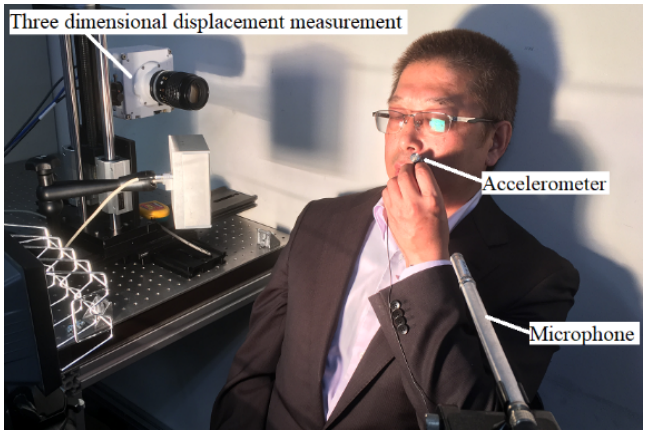


FIGURE 1. Sound recording

TABLE 1. Experimental setup

Equipment	Model name
Software	Istra4D, Dantec Dynamics
Measurement	Q-400, Dantec Dynamics
Camera	VEO410L, Phantom
Lens	NIKKOR 105mm, Nikon
Lighting	Halogen & LED lighting
Microphone	MI-1431, Ono Sokki
Accelerometer	NP-2110, Ono Sokki
Amplifier	SR-2200, Ono Sokki

use of environment and equipment for three-dimensional displacement measurement. A participant who is a healthy male vocalizes a speech while sitting on the chairs, and seven conditions are selected from the vocabulary, including five Japanese vowels, one syllabic nasal /N/, and a quiet silence. Under the experimental conditions shown in Table 1, the microphone is installed at a distance of about 30 cm from the mouth, and the accelerometer is placed on top, where it is employed as the measurement location [6]. The cameras for three-dimensional displacement measurement were focused on the opposite side of the upper lip for the accelerometer. It is necessary to measure the 4th or 5th formant, such as from about 4 to 5 kHz or more; however, the ability to sample frequency has a relationship to trade-off between the frequency and the conditions of measurement, such as luminance and memory for capturing images. Considering this relationship, 10 kHz sampling was used in this experiment with 0.2 s recording as a maximum performance of this experiment.

3.2. **Result and discussion.** Figures 2 and 3 show speech and body-conducted speech. In the speech, the fundamental frequency and the 1st to 4th formants were observed on the spectrum and its envelope, and the frequency characteristic of body-conducted speech was measured with up to approximately 1 kHz frequency. Spectrum and its envelopes illustrate with Power Spectral Density Function (PSD), Liner Predictive Coefficient (LPC), and Cestrum analysis. PSD is conventional spectral analysis, and LPC and Cepstrum analysis are methods for describing the Spectrum envelope. The differences become a factor that reduces sound quality and leads to the dismissal of the 3rd and higher formants. In general, vocabulary can be recognized from the 1st and 2nd formats, so there is no problem with hearing vocabulary in spoken words and sentences of speech. Figure 4 shows the x and y axes of skin-propagated speech measured on the upper lip. A vowel is a vocalized sound with a vocal cord, for which a repetition of the fundamental wave is seen in the waveforms

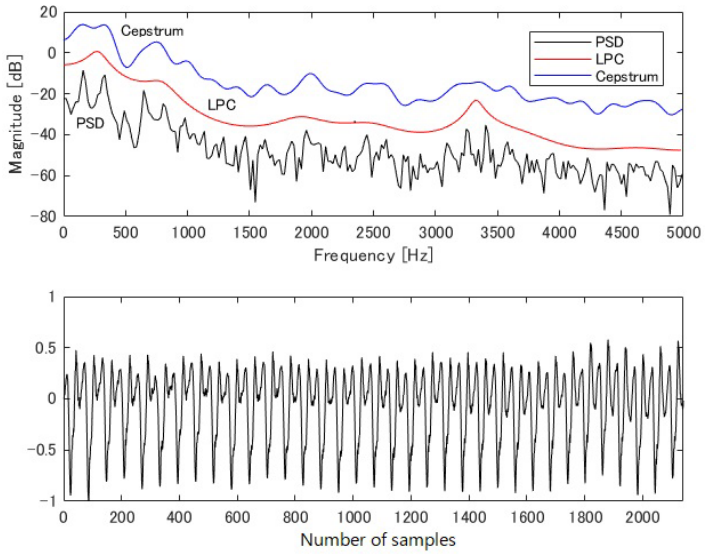


FIGURE 2. Speech

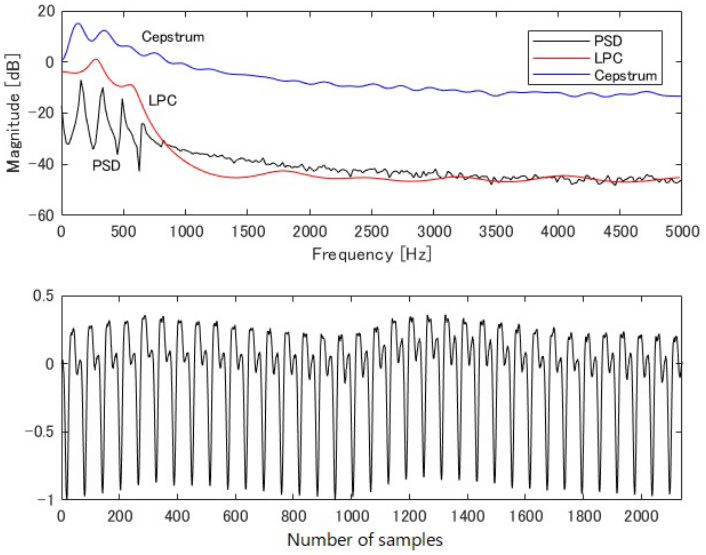
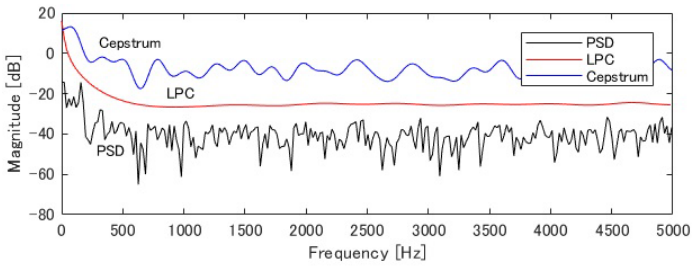
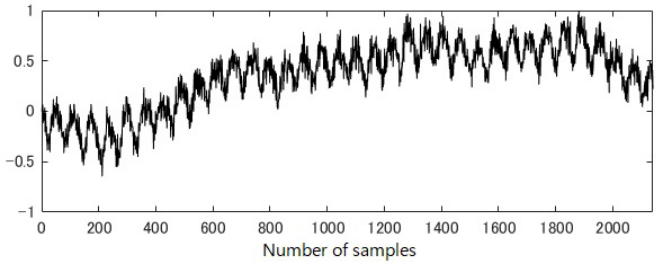


FIGURE 3. BCS

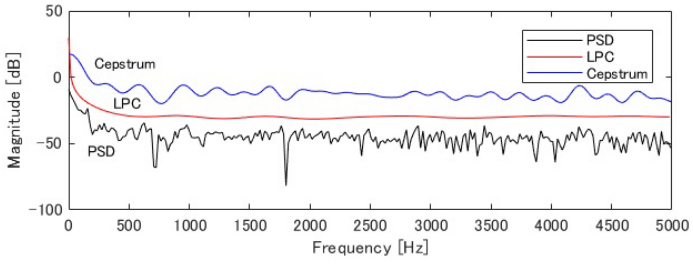
of both axes on the spectrum. Attenuation of approximately 10 dB occurs up to nearly 500 Hz rapidly after the fundamental frequency, and the attenuation of the gain can be confirmed in the order of speech, body-conducted speech, and propagation through the skin. In addition, as the magnitude varies depending on the axis, it can be seen that there is a bias in the direction of vibration, respectively. Based on these results, the following two outcomes are considered from this experiment. This bias means the expected dependence of the axis on human skin, so this is one of the important outcomes indicating a necessity for three-dimensional displacement measurement. In addition, it is believed that a transfer function can be obtained that converts acceleration and displacement dimensions at the same measurement location, such as the upper lip because there were only differences in size between the displacement and the accelerometer during a signal measurement.

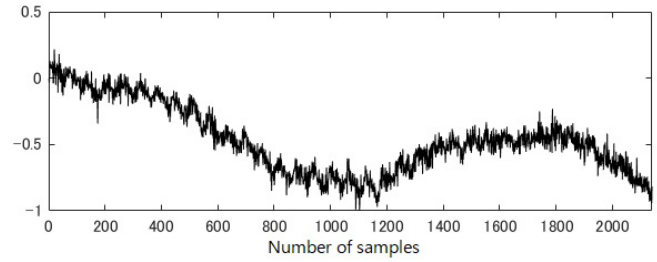
4. **Experiment 2: Vibration Strength Depending on Axis Direction.** A human face is composed of different materials with different characteristics, such as bones and muscles, and it has a complex structure. Therefore, it is assumed that there is a bias in the direction and intensity of the transmitted sound. Here, it is confirmed that there is a vibration bias along the x and y axes.





(a) x axis





(b) y axis

FIGURE 4. Skin propagated speech

4.1. **Experiment setup.** Five Japanese vowels recorded using three-dimensional displacement measurements were also used in this experiment. In general, speech sounds have separate characteristics, namely the vocal code characteristics and articulation information. Vocal code characteristics are a source of sound similar to a pulsed sound, and articulation information is a frequency characteristic composed of oral and nasal cavities. The experimental goal is to measure the vibration strength depending on axis directions, so the spectrum of each axis is calculated from the spectral envelope, such as articulation information. To express the characteristics of vocal and nasal cavities, Linear Predictive Coefficient (LPC) [6] describes spectral envelope, which uses eight dimension coefficients and is analyzed.

4.2. **Result and discussion.** Figure 5 shows the spectral envelope of the vowel /a/ of skin vibration sound calculated using LPC analysis. This sound is a displacement which contains an offset in the form of DC, so the displacement changes to the sound velocity in a differential calculation. Then, the spectral magnitude is calculated as a result of the x axis meaning the vertical direction of a face, and the y axis means its horizontal direction.

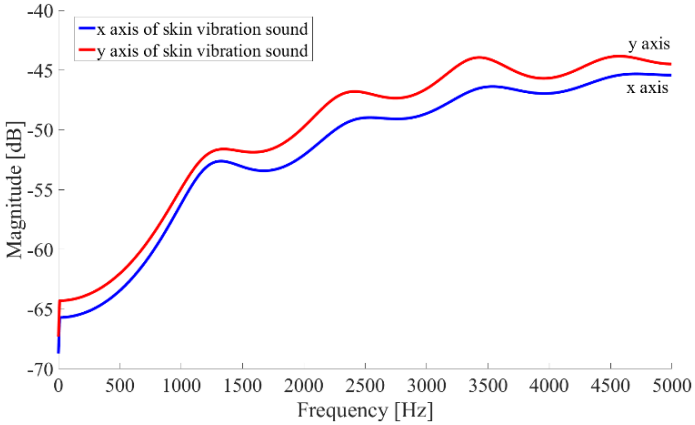


FIGURE 5. Spectrum envelope of the vowel /a/

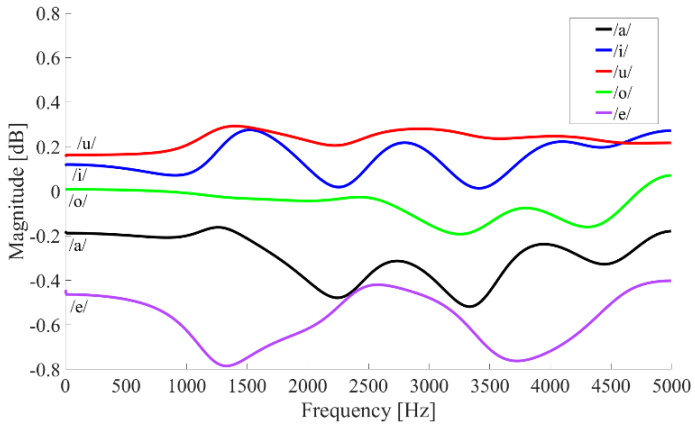


FIGURE 6. Magnitude bias dependent on axis

Figure 6 shows the dependence of vibration strength on axis directions. The differences of y and x magnitudes were calculated. There is little difference in Figure 5, but in Figure 6 it is easy to see the difference. In Figure 6, a positive value indicates the strength of y axis, and a negative value indicates the strength of x axis. Thus, the vowel /a/ is confirmed by a bias towards y axis. The other four vowels are also calculated in a similar way. As a result, it was confirmed that the vibrations were biased in the vertical axis direction for /a/ and /e/ and in the horizontal axis direction for /i/ and /u/, respectively. For /o/, there is almost no dependence of the direction on the axis in the low-frequency band, and it can be mentioned that the vibration is slightly biased in the vertical axis direction at around 3 kHz and 4.5 kHz. In /i/ and /u/, the dependence of the direction in the horizontal direction was large at approximately 1.5 kHz, 2.8 kHz, and 4 kHz. In /a/, the vibration bias in the vertical axis direction was large at nearly 2.2 kHz, 3.3 kHz, and 4.5 kHz. Regarding /e/, the dependence of the direction along the vertical axis was large at approximately 1.3 kHz and 3.7 kHz. It is believed that the cause of this difference is related to the position of the tongue and the shape of the lips during the utterance. Among these results, /i/, /e/ and the other three are classified as three patterns of skin vibration.

5. **Experiment 3: Measurement for Fundamental Frequency.** We investigated to estimate the fundamental frequency components that could be measured fundamental frequency from recorded skin propagation speech. Figure 7 shows the frequency characteristics of x axis skin vibration for vowels /i/ and syllabic nasal /N/. The average power spectrum was obtained with a frame length of 51.2 ms and an overlap of 25.6 ms. A

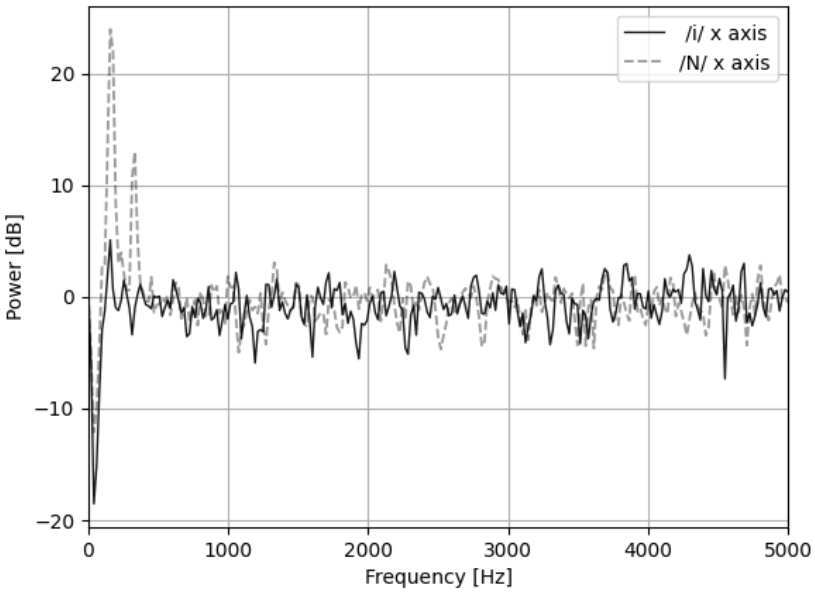


FIGURE 7. Spectrum of /i/ and /N/ on x axis

high-pass filter with a passband of 100 Hz or higher was having for the three-dimensional displacement measurement, considering the natural quivers of the human body. The peaks of the fundamental frequency components of both axes were confirmed at the same fundamental frequency, and in particular, the second harmonic component of /N/ was also clearly identified; however, the problem with /i/ is its low SNR. This difference can be attributed to the fact that /N/ is a closed-mouth vocalization, whereas /i/ is an open-mouth vocalization. Thus, the sound energy is converted to skin vibration energy. However, in all cases, the noise component is large relative to the fundamental frequency, making it difficult to extract the fundamental frequency. Therefore, it is necessary to apply noise reduction methods such as spectral subtraction and Wiener filtering to skin propagation speech.

5.1. **Experimental setup.** Based on the results of frequency analysis, we assumed that skin vibration is a signal in which noise is suspended in speech, and we expected to extract the fundamental frequency if noise reduction is applied to skin propagated speech. Skin propagated sound $s(t)$ is expressed by the following Equation (1).

$$s(t) = x(t) + n(t) \tag{1}$$

Here, $x(t)$ is skin propagated speech during vocalization and $n(t)$ is the skin propagated speech during silence in this research. In this experiment, we estimate the speech signal from the skin propagated speech using Spectral Subtraction method (SS method) and Wiener filter method for x axis skin vibration signals of /N/ and /i/, and extract the fundamental frequency from the estimated speech using Harvest method, an effective extraction method for low SNR signals, in WORLD [8].

Spectral subtraction method [9]

SS method estimates a speech by subtracting the average power spectrum of noise from power spectrum of noisy speech, which is expressed by the following Equation (2).

$$P_s(f) = P_x(f) - \mu\overline{P_n}(f) \tag{2}$$

$P_s(f)$, $P_x(f)$, and $P_n(f)$ are the estimated speech, noisy speech, and noise spectrum, respectively, and μ is scaling coefficient. We attempt to reduce this noise by updating these and repeating the process.

Wiener filtering method [10]

The Wiener filter method estimates retrieval speech as a transfer characteristic $H(f)$, which can be obtained by the following Equations (3) and (4).

$$H(f) = \frac{P_x(f)}{P_x(f) + \mu\overline{P_n}(f)} \tag{3}$$

$$P_s(f) = H(f)P_x(f) \tag{4}$$

$P_s(f)$, $P_x(f)$, and $P_n(f)$ are retrieval speech, noisy speech with noise, and noise spectrum, respectively, and μ is scaling coefficient. Noise reduction is attempted by repeating these steps at several times while updating the transfer functions.

5.2. **Result and discussion.** Figure 8 shows the results of fundamental frequency to extract for /N/ and /i/. Both methods reduced in error of estimations between speech and these methods, and in particular, the Wiener filter method extracted the fundamental frequencies precisely. However, fundamental frequency of /i/ could not be extracted by both of the methods. This is due to the lower SNR of /i/ compared to /N/, and further study is needed.

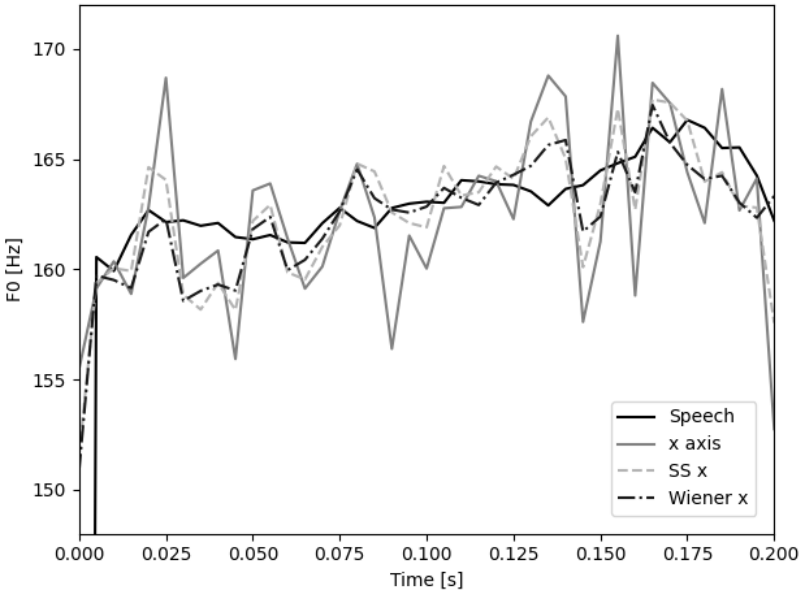


FIGURE 8. Fundamental frequency trajectories of /N/

6. **Conclusion and Future Work.** In order to confirm the possibility of skin measurement using a three-dimensional displacement measurement, speech, body-conducted speech, and skin propagation on the upper lip were measured. Attenuation of approximately 10 dB was observed up to nearly 500 Hz on the skin surface via body-conducted speech. In particular, there are three findings: the direction and attenuation depend on each axis, a transfer function can be obtained that converts between the dimensions of acceleration and displacement at the same measurement location on the lip, and the vibration bias can be classified into three groups. In estimation of fundamental frequency, Wiener filter was used to estimate a fundamental frequency close to a trajectory of speech.

In future, we will discuss ways to estimate the transfer functions that convert the displacement and acceleration; further, the measurement using three-dimensional displacements with the added z axis will be performed as a truly three-dimensional measurement. In addition, the vibration using three dimensions for more precise analysis will be attempted to visualize detailed skin propagation and vibration.

Acknowledgment. We would like to thank Mr. Y. Futawatari, Dantec Dynamics, for coordinating this experiment. We would also like to thank Mr. K. Tomioka and Mr. M. Sato, CORNES Technology Limited, for the opportunity to measure three-dimensional displacements, and also thank Mr. G. Kuwabara, Photron Limited, for supporting the experiment with high-speed camera.

REFERENCES

[1] H. Tabatabai, D. E. Oliver, J. W. Rohrbaugh and C. Papadopoulos, Novel applications of laser Doppler vibration measurements to medical imaging, *Sensing and Imaging: An International Journal*, vol.14, pp.13-28, 2013.

[2] T. Kitamura, Verification of reproducibility of measurements of skin vibration during singing by scanning laser-Doppler vibrometer, *The Japan Journal of Logopedics and Phoniatrics*, vol.55, no.2, pp.167-172, 2014 (in Japanese).

[3] Y. Avargel and I. Cohen, Speech measurements using a laser Doppler vibrometer sensor: Application to speech enhancement, *Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pp.109-114, 2011.

[4] S. Lee, J. Kim, I. Yun, G. Y. Bae, D. Kim, S. Park, I. M. Yi, W. Moon, Y. Chung and K. Cho, An ultrathin conformable vibration-responsive electronic skin for quantitative vocal recognition, *Nature Communications*, vol.10, 2468, 2019.

[5] Dantec Dynamics, *Vibration and Modal Shape Analysis*, Application Note, 2016.

[6] S. Ishimitsu et al., A noise-robust speech recognition system making use of body conducted signals, *Acoustical Science and Technology*, vol.2, no.25, pp.166-169, 2005.

[7] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America*, vol.87, no.4, pp.1738-1752, 1990.

[8] M. Morise, F. Yokomori and K. Ozawa, WORLD: A vocoder-based high-quality speech synthesis system for real-time applications, *IEICE Transactions on Information and Systems*, vol.E99-D, no.7, pp.1877-1884, 2016.

[9] M. Morise, T. Irino and H. Kawahara, Error evaluation of impulse response estimation by cross spectral method using speech signal, *Journal of IEICE*, vol.J90-A, no.7, pp.559-566, 2007.

[10] D. Li and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*, Marcel Dekker Inc., New York, 2003.