LOGARITHMIC TEMPORAL PROCESSING APPLIED TO ACCURATE EMPIRICAL TRANSFER FUNCTION MEASUREMENTS IN VOCAL SOUND PROPAGATION

Masanori Morise, Toshio Irino, and Hideki Kawahara

Faculty of Systems Engineering, Wakayama University
930 Sakaedani, 640–8510, Wakayama, Japan
phone: +81-73-457-8525 fax: +81-73-457-8525 , email: s055068@sys.wakayama-u.ac.jp
web: http://media.sys.wakayama-u.ac.jp/AuditoryMediaLab

ABSTRACT

A new procedure to improve accuracy in an empirical transfer function measurement method is proposed for investigating speech sound propagation. In our previous work, vowel dependent behavior of empirical transfer functions from a lip reference point to observation points around a speaker's head was found. The accuracy of the method was also evaluated by using references obtained using a HATS and an M-sequence that revealed significant accuracy degradations in higher frequency range due to low speech energy. The proposed method solves this problem by introducing logarithmic temporal manipulation and low-pass filtering. The proposed method was tested using 186 vocalizations of sustained Japanese vowels. Test results indicated that the proposed method reduced standard deviations to 80% in gain estimation, 33.8% in weighted group delay estimation, and 20% in duration estimation, respectively, in frequency regions higher than 10 kHz. Detailed implementation aspects are also discussed.

1. INTRODUCTION

There are two basic strategies in sound reproduction. One reproduces the surrounding sound field into the listener's environment. The other reproduces sound radiation from the sound source into the listener's (real and/or virtual) environment. This article provides a new method for measuring sound radiation when the source is a human voice.

Historically, the radiation patterns of speech sounds have been investigated using artificial mannequins [1]. They have also been standardized for telephone system measurements as head and torso simulators (HATS) [2, 3]. However, this does not necessarily suggest that radiation patterns are understood enough for realistic simulations. Several factors must be taken into account when considering the sound radiation patterns of human speech. Suzuki reported significant sound radiations from various parts of a speaker's body [4]. Variations of mouth opening while speaking also may introduce radiation pattern variations due to changes in deflection. These factors make it inevitable to use speakers' own voices as test signals for measurements. This requirement introduces difficulties in reliable measurements.

In our previous study, a cross-spectral method was applied to investigate such vowel dependent changes in empirical transfer functions that clearly indicated that such variations do exist [5]. The measurements, however, failed to provide reliable results in higher frequency regions, for example, 8 kHz or higher due to low signal to noise ratio.

This article provides a systematic and efficient method for improving this signal to noise ratio to yield reliable results by applying logarithmic temporal manipulation and filtering in a manipulated domain [6, 7]. First, the general setup of voice empirical transfer function measurements and discussions on issues in cross-spectral methods are briefly introduced. Following the principles of the operations of logarithmic manipulation and filtering, detailed descriptions and test results using one male speaker are presented that illustrate the effectiveness of the proposed method in an anechoic chamber.[1]

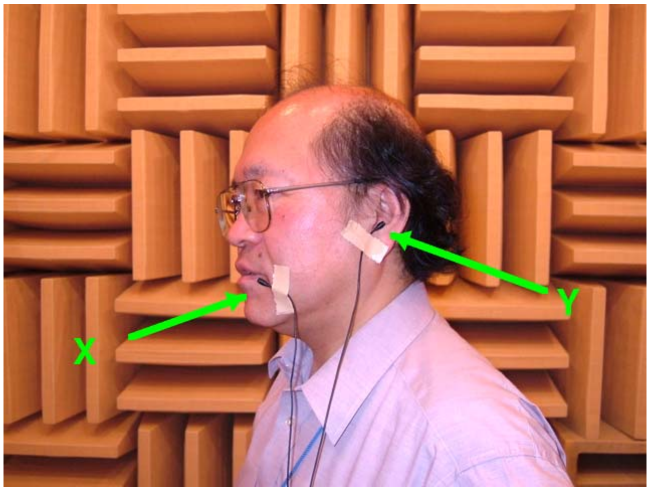

Figure 1: Experimental setup showing subject and microphone placement

2. OUTLINE OF MEASUREMENTS USING A CROSS-SPECTRAL METHOD

Standard methods for measuring transfer functions employ specially designed test signals, such as M-sequence and Time-Stretched pulses (TSP). These signals have a constant energy distribution in each frequency. In contrast to these signals, a test signal, which is a speaker's own voice, has frequency dependent energy distribution. A cross-spectrum based method was adopted to handle this situation.

[1] This is an important extension to our previous reports [6, 7] where measurements were conducted in a sound-proof room with fewer voicing repetitions.

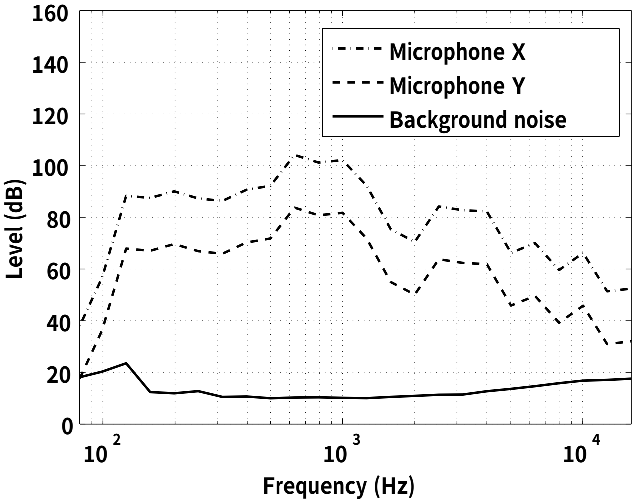

Figure 2: Background noise, microphone X, and microphone Y. Level is calculated from output of one-third octave band filters.

2.1 Experimental setup and recording conditions

Figure 1 shows the placement of microphones for recording speech sounds. Reference microphone (represented as “X” in the figure) was placed off axis to prevent pop noise due to expiration and blowing. The microphone for the measuring point (represented as “Y” in the figure) was placed at the entrance of the speaker’s left ear canal.

Recordings were conducted in an anechoic chamber at ATR. Small omni-drectional condensor microphones (DPA 4060-BM) were used for measurements to minimize sound field disturbance. Two channel microphone output was preamplified by microphone amplifiers (TASCAM MX-4) and recorded using an audio interface (RME Hammerfall) to a PC. The sampling rate and resolution were 44100 Hz and 24 bit. The one Japanese male subject who participated in the experiment was asked to produce sustained Japanese vowel /a/ with a constant fundamental frequency (F0) and a regularly varying F0 (roving F0 condition).[2] The total number of voicing segments was 186. The average duration of segments was about eight seconds.

Figure 2 shows one-third octave band levels for two microphone outputs of a voiced segment and the background noise. Note that the signal to noise ratios for microphone Y in the higher frequency bands (namely, 10 kHz or more) are 20 dB or less and that virtually no voice energy is observed in the 80 Hz frequency band of the Y microphone output. These low signal to noise ratios result in estimation errors described in the next section.

2.2 Cross-spectral estimation of transfer functions

An empirical transfer function $T_k(\omega)$ of the k-th voicing segment is calculated using the following standard cross-spectral method:

$$T_k(\omega) = \frac{\langle Y_k(\omega)X_k^*(\omega)\rangle}{\langle X_k(\omega)X_k^*(\omega)\rangle}, \qquad (1)$$

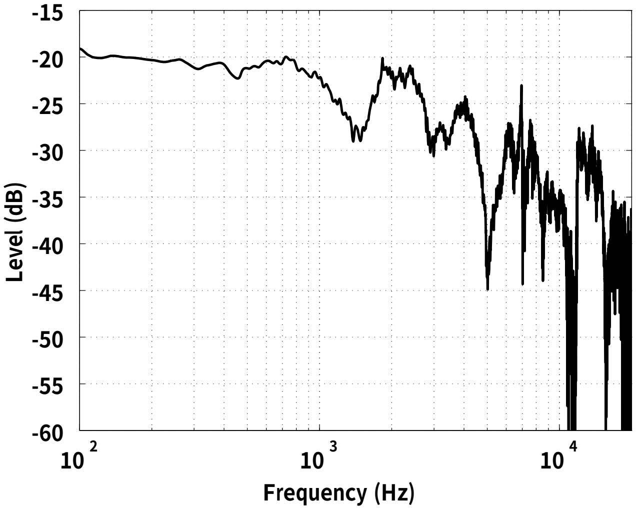

Figure 3: Estimated amplitude transfer function

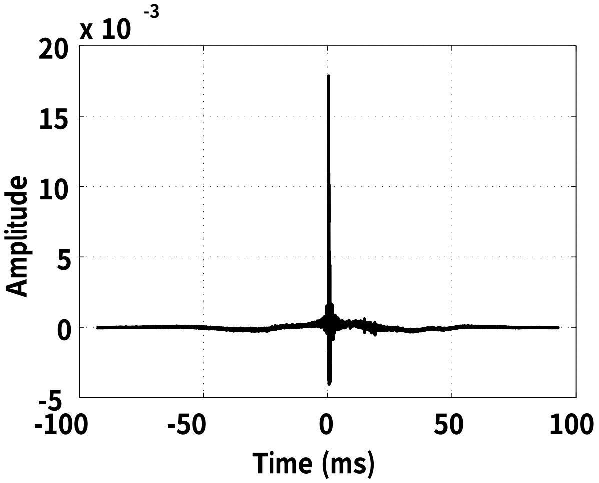

Figure 4: Estimated impulse response from transfer function in Figure 3

where angle brackets represent ensemble average, $X_k(\omega)$ and $Y_k(\omega)$ represent the complex spectra of reference and measuring signals, respectively, and $*$ represents a complex conjugate. In this experiment, ensemble average was replaced by sample average using a sliding Blackman window (8192 samples in length) with 90% overlapping (819 samples for frame shift).

Figure 3 shows an example of an estimated amplitude transfer function $|T_k(\omega)|$. Note that the line width of the plot looks thickened in the higher frequency region, for example, 4 kHz or more. This thickening is caused by abrupt changes in amplitude and is salient, especially in regions higher than 10 kHz. Amplitude also fluctuates in the lower frequency region (less than 100 Hz, in this case). These observations are consistent with low signal to noise ratio in the frequency bands shown in Figure 2.

Figure 4 shows impulse response calculated from the empirical transfer function shown in Figure 3. A slowly moving

[2] The roving F0 condition was designed to prevent spectral zeros between harmonic components in long-term power spectra [5].

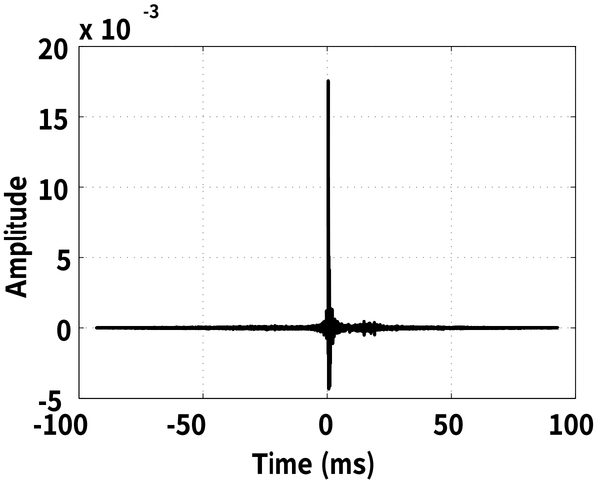

Figure 5: Estimated impulse response from transfer function using frequency domain constraint

level in the response corresponds to fluctuations in the lower frequency region, and distributed noisy variations correspond to abrupt changes in the higher frequency region.

2.2.1 Simple post-processing for fixing inaccuracy

These defects result from low signal to noise ratio, in other words, a shortage of information. Several apriori constraints can be used to supply the missing information.

Frequency domain constraint: Since the distance between microphones is less than 30 cm, the amplitude in frequency response stays virtually constant in lower frequency regions less than 100 Hz. Phase shift at frequency f, which resulted from propagation delay τ_p, is represented as $2\pi f \tau_p$ in the lower frequency region. These yield the following equation for substituting the lower portion of the measured transfer function:

$$T_k(\omega) = |T_k(\omega_0)| e^{j\phi(\omega_0)\frac{\omega}{\omega_0}}, \qquad (2)$$

where $f_0 = \omega_0/2\pi$ is the lowest frequency that provides a reliable estimate of the transfer function. In this measurement, 125 Hz is selected based on the one-third octave signal to noise ratio shown in Figure 2. Figure 5 shows the impulse response calculated by using Eq. 2.

Time domain constraint: Since the primary radiating sound source is the subject's mouth, responses coming later than 7 ms from the dominant peak are reflections from objects other than the subject's own body. Responses preceding 1 ms to the dominant peak are also due to background noise. Removing those components using a sigmoidal truncating window yields the estimated amplitude response shown in Figure 6.

2.3 Reproducibility of measurements

The reproducibility of measurements was tested by analyzing 186 recorded segments. Figure 7 shows the averaged amplitude frequency characteristics (the heavy line), and the standard deviations (±2σ) of the measurements (thin lines). Larger standard deviations in the higher frequency region suggests that other sources of variations in the measurements still remain.

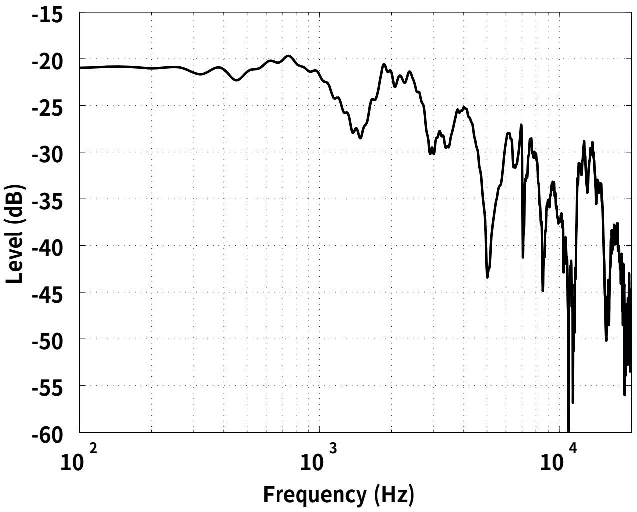

Figure 6: Estimated amplitude transfer function using both frequency and time domain constraints

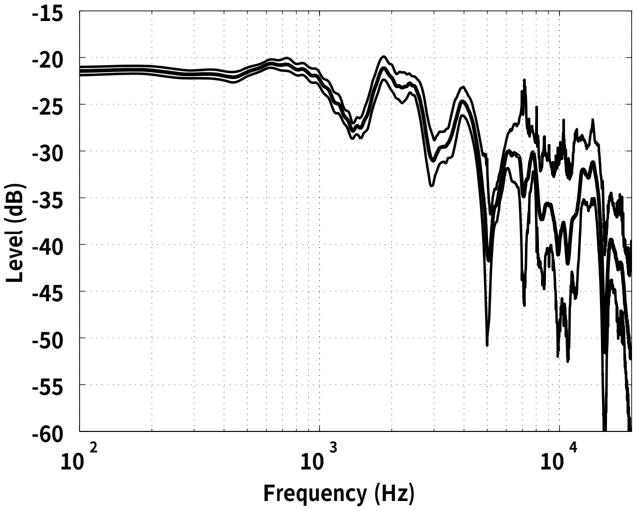

Figure 7: Averaged amplitude frequency characteristics (the heavy line) and standard deviations(±2σ) of measurements (thin lines)

The reproducibility of temporal aspects is evaluated using weighted group delay. Because averaged time $\langle t \rangle$ and duration σ_t of response $s(t)$ are defined in the time domain, they are also represented using group delay $\tau_g(\omega) = -\psi'(\omega)$ and amplitude spectrum $B(\omega) = |S(\omega)|$, as shown in the following equations [8]:

$$\begin{aligned} \langle t \rangle &= -\int \psi'(\omega)|S(\omega)|^2 d\omega & (3) \\ \sigma_t^2 &= \int \left(\frac{B'(\omega)}{B(\omega)}\right)^2 B^2(\omega) d\omega & \\ &\quad + \int (\psi'(\omega) + \langle t \rangle)^2 B^2(\omega) d\omega & (4) \\ S(\omega) &= |S(\omega)| e^{j\psi(\omega)} = B(\omega) e^{j\psi(\omega)}, & \end{aligned}$$

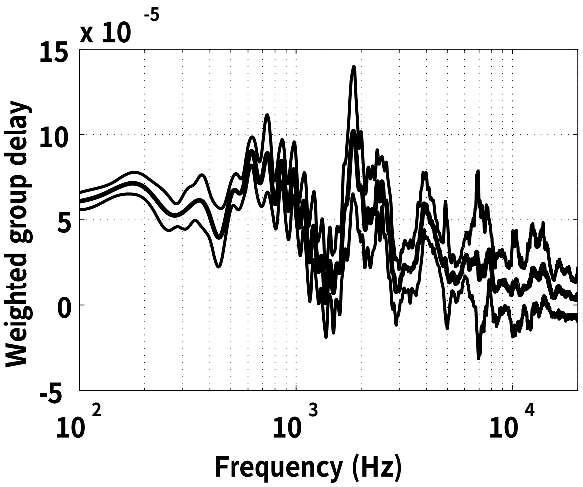

Figure 8: Average of weighted group delay (the heavy line) and standard deviation (±2σ, thin lines)

where $'$ represents derivative in terms of ω. For the simplification of the equations' appearances, response power is assumed to be normalized [8].

Figure 8 shows the average of weighted group delay $\tau_g(\omega)B^2(\omega)$ (the heavy line) and standard deviations (±2σ) of the measurements (thin lines). Deviations are evenly distributed in all frequency regions but are slightly larger in the higher frequency region.

There are two major source of these deviations. One is the variations of the positioning of microphones because the reference microphone was attached to the facial tissue close to the mouth that can move while voicing. The other source is background noise due to ambient and electronic circuit noise. The goal of this article is to reduce the effects of the second source.

3. LOGARITHMIC TEMPORAL MANIPULATION AND FILTERING

If we have access to prior knowledge about the time-frequency region where causal response components are dominant, we can improve the accuracy of the estimated impulse response using components residing within such a reliable region. The causal impulse response consists of directly propagated components, components due to the diffraction of the speaker's body, reflections from body parts, and other reflections. The desired response should consist of all causal components but without other reflections. Therefore, the reliable region is determined based on the ratio between the desired response and background noise.

The following simplifications were introduced to outline the reliable region: a) The amount of diffraction is assumed to be inversely proportional to component frequency; b) Reflections from a speaker's body, captured by the measuring microphones, are diffractions at the reflecting point and are integrated all over the speaker's body. These assumptions suggest that the duration of a frequency component is also inversely proportional to its frequency. This time-frequency selection and reconstruction of the impulse response can be implemented by employing a wavelet transform or other joint time-frequency representations. However, another efficient implementation exists that uses logarithmic temporal manipulation and filtering [6].

3.1 Designing temporal manipulation

Assume that the boundary of the desired region is defined by function $b(t)$. By defining a new temporal axis $\tau(t)$ using the following mapping, apparent instantaneous frequency of the boundary defining function stays at constant value f_L on this new temporal axis:

$$\tau(t) = \int_{t_0}^{t} \frac{b(\lambda)}{f_L} d\lambda. \quad (5)$$

When the boundary function is inversely proportional to the distance from the origin of the time axis, the mapping function yielded by Equation 5 is logarithmic. Then filtering out components higher than f_L by low-pass filtering on new temporal axis $\tau(t)$, inverse mapping back to the original temporal axis yields the desired time-frequency region selection. This is the basic idea of the proposed logarithmic temporal manipulation and filtering.

3.1.1 *Implementation details*

An issue must be considered for implementing this basic idea in the current measurements: the multipath problem. Voiced sound propagation delays including bone conduction differ depending on propagation paths. Assume that they are distributed in $(0, t_s)$ as the first order approximation. The next assumption is that the envelope of responses at each frequency has exponential decay with damping rate ζ. Then, by assuming an S/N threshold to r, the new time axis $\tau(t)$ for implementing the desired time-frequency selection is yielded as follows:

$$\begin{aligned} \tau(t) &= \frac{\log(r)}{2\pi\zeta f_L}\left(\log(t-t_s) - \log(t_1)\right) + \frac{f_s(t_1+t_s)}{2 f_L} \quad (6) \\ & \text{where} \quad t_1 = \frac{\log(r)}{\pi f_s \zeta}, \end{aligned}$$

where f_s represents sampling frequency and t_s represents the maximum propagation lag difference. This equation holds for $t > t_1 + t_s$. For $0 < t < t_1 + t_s$, it is reasonable to set an effective cut-off to the Nyquist frequency.

Inverse function $t(\tau)$ of this warping is represented by the following equation.

$$\begin{aligned} t(\tau) &= \begin{cases} \frac{2 f_L \tau}{f_s} & \tau < f_s(t_1+t_s)/f_L \\ t_1 \exp\left(\frac{\tau - C_2}{C_1}\right) + t_s & \tau \geq f_s(t_1+t_s)/f_L \end{cases} \quad (7) \\ & \text{where} \quad C_1 = \frac{\log(r)}{2\pi\zeta f_L}, \quad C_2 = \frac{f_s(t_1+t_s)}{2 f_L}. \end{aligned}$$

4. EVALUATIONS

The proposed method was evaluated using the recorded 186 voicing segments. Figure 9 shows average amplitude frequency characteristics (the heavy line) and standard deviations (±2σ, thin lines). Figure 10 shows the average of weighted group delay $\tau_g(\omega)B^2(\omega)$ (the heavy line) and standard deviation (±2σ, thin lines). The upper plots in Figures 9 and 10 are identical to Figures 7 and 8, respectively. The lower plot shows the results of the proposed method. The

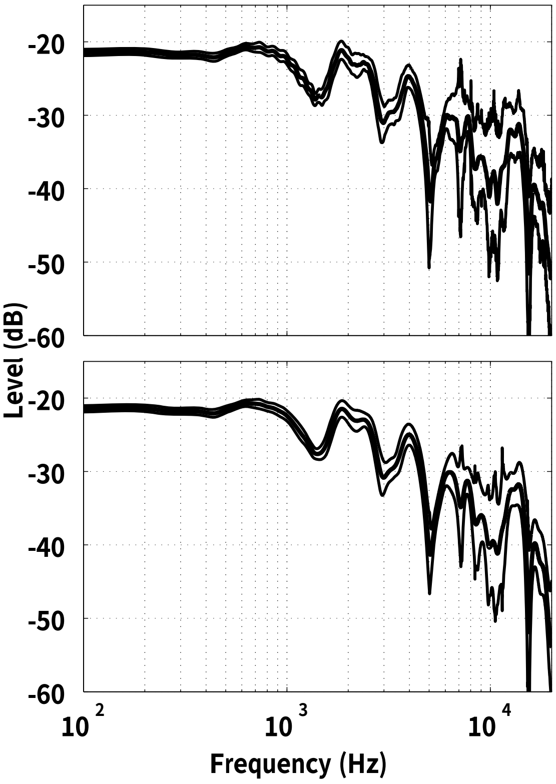

Figure 9: Average amplitude frequency characteristics (the heavy line) and standard deviations (±2σ, thin lines). Upper plot is identical to Figure 7. Lower plot shows results of proposed method.

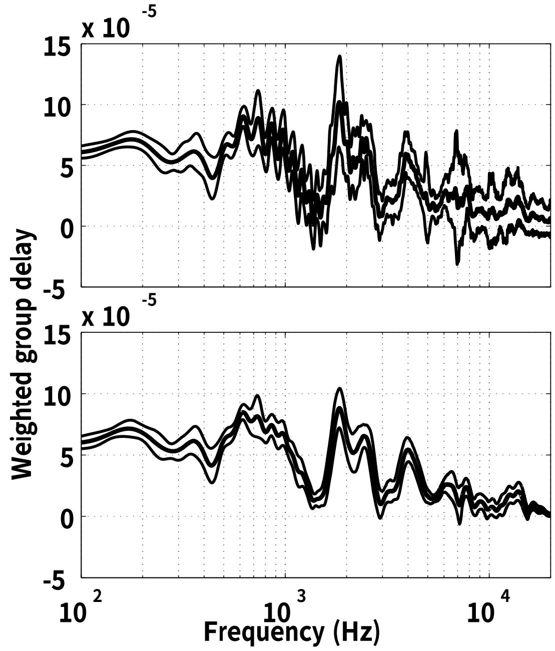

Figure 10: Average of weighted group delay (the heavy line) and standard deviations (±2σ, thin lines). Upper plot is identical to Figure 8. Lower plot shows results of the proposed method.

test results indicated that the proposed method reduced standard deviations to 80% in gain estimation, 33.8% in weighted group delay estimation, and 20% in duration estimation, respectively, in frequency region higher than 10 kHz. Although accuracy improvement in amplitude response estimation in terms of reproducibility was not substantial, significant improvement was achieved in the temporal domain parameters. Note that smoothness in the estimated amplitude responses, which resulted from the proposed method, is crucially important in designing compensating inverse filters.

5. CONCLUSION

A new method for voice radiation measurement based on logarithmic temporal manipulation was proposed that features post-processing for improving the accuracy of measurements. The proposed method was evaluated by using 186 voiced segments produced by a Japanese male subject in an anechoic chamber, and its effectiveness was revealed especially in the temporal aspects of the estimated response. Voice radiation pattern measurements using the proposed method are currently under preparation using a computer-controlled turntable and multiple microphones and will be reported elsewhere.

Acknowledgments

This research was supported partly by grant in aids in scientific research (17650054) and the CrestMuse project by JST.

REFERENCES

[1] J.L. Flanagan, “Analog measurement of sound radiation from the mouth,” J. Acoust. Soc. Am., vol. 32, no.12, pp. 1613–1620, 1960.

[2] ITU-T series p. 51, Telephone transmission quality, Objective measuring apparatus, Artificial Mouth.

[3] ITU-T series p. 58, Telephone transmission quality, Objective measuring apparatus, Head and torso simulator for telephonometry.

[4] H. Suzuki, Dang. J, and Nakai. T, “Measurements of sound rediation and vibration at lip, nostril and laryingeal area and simulation of acoustic leakage between nasal cavity and oral cavity,” Journal of IEICE, vol. J74-A, no.12, pp. 1705–1714, 1991. (In Japanese)

[5] M. Nukina, H. Kawahara, “Cross spectral measurement of head related speech transfer functions using speaker’s own voice,” First Pan-American/Iberian meeting on acoustics, 2-6 December 2002, Cancun, (J. Acoust. Soc. Am. 112, p. 2323 (2002)).

[6] M. Morise, H. Kawahara, “Loudspeaker equalization based on multi-location observation with reliable time-frequency region selection and its evaluation using sound propagation measurement,” Proc. EUSIPCO’2004 Vienna, pp. 1995-1998, 2004.

[7] M. Morise, T. Irino, H. Kawahara, “Accuracy improvement in speech sound propagation measurement using logarithmic temporal manipulation,” IEICE technical report, EA2005-64, Oct, 2005. (In Japanese)

[8] L. Cohen, “Time-frequency analysis,” Prentice Hall, Englewood Cliffs, NJ, 1995.