Sound & Vibration
DOI:10.32604/sv.2020.08119
Article

Tech Science Press

Effect of Input Waveform to Vibration Speaker on Sound Quality of Electric Artificial Voice

T. Asakura* **and K. Shindo**

Tokyo University of Science, Chiba, 278-8510, Japan
*Corresponding Author: T. Asakura. Email: t_asakura@rs.tus.ac.jp
Received: 29 July 2019; Accepted: 10 February 2020

Abstract: A number of studies have been conducted on the improvement of the sound quality of electrical artificial laryngeal speech, the speech produced has been difficult to hear compared to a natural voice. For this reason, it is necessary to effectively improve the frequency characteristics of the input signal. In the present study, to improve the sound quality of vocalization using an electrical artificial larynx, first, a comparison of the frequency characteristics between the real and artificial voices was conducted, and three filters that can make the frequency characteristics of the artificial voice closer to those of a real voice were generated. Then, the influence of these filters on the quality of the artificial voice was investigated via physical measurement and a subjective evaluation experiment targeted at Japanese five vowels. It was found that the intelligibility of artificial /a/ and /o/ sounds was improved, whereas little improvement was observed in the case of /i/, /u/, and /e/. The obtained results confirmed the effect of optimizing the input signal into the vibration speaker and indicated areas for further improvement.

Keywords: Electriolaryngeal voice; intelligibility; naturalness; loudness

1 Introduction

Speech is one of the most common communication tools and is used to express feelings and convey opinions. However, individuals who have had their larynx removed by surgery for laryngeal cancer cannot speak with their own vocal cords, which not only reduces their communication ability, but also causes mental distress as a result of not being able to speak. After such surgery, since the substitute trachea is opened in the throat during a laryngectomy to secure the airway separately from the oral cavity, the patient loses the function of the vocal cords, which work based on the airflow from the lungs. As such, the electric artificial larynx was developed as an alternative vocalization method for laryngectomy patients.

Using the electric artificial larynx, laryngectomy patients can produce sounds by inputting a signal that simulates vocal cord vibration, because, in most cases, the articulatory organs, such as the oral cavity and the tongue, are used in producing sound. However, since the input signal applied to the electric artificial larynx is monotonous and mechanical, speech produced by the electric artificial larynx is difficult to understand.

A number of studies have been conducted on electrical artificial laryngeal speech. The modern electrolarynx was first proposed by Barney [1]. Holley et al. performed a comparative study on the

intelligibility of esophageal, electrolaryngeal, and normal speech in quiet and noisy fields [2]. Weiss et al. investigated the fundamental acoustic characteristics of the speech produced with an electronic artificial larynx targeting normal-speaking adult males who were trained to use an electrolarynx [3]. In a follow-up study [4], they also performed a closed-set word discrimination test, and correct identification of words was discussed.

In order to improve the quality of the electrolarynx speech performance, various investigations have been performed. Liu et al. applied a spectral subtraction method to efficiently enhance the artificial laryngeal voice [5,6]. They have also applied the spectral subtraction method to the enhancement of electrolarynx speech [7]. Basha et al. investigated the real-time enhancement of electrolarynx speech by using the spectral subtraction method [8]. The feasibility of using a motion sensor to replace a conventional electrolarynx user interface equipped with on/off and pitch frequency control was explored by Matsui et al. [9]. Tanaka et al. presented an electrolaryngeal speech enhancement method capable of significantly improving naturalness of the electrolaryngeal speech [10,11]. When exciting the neck by a vibrator, it is important that the vibration is efficiently transmitted to the vocal tract through the neck tissue. In order to aid in the design of improved neck-type electrolarynx devices, Meltzner measured the neck frequency response function [12]. In order to efficiently measure such transmission characteristics through neck tissue, a simple method using a reflectionless tube was proposed [13].

For the sake of improvement of the usability of an electrolarynx, the following investigations have been performed. Ifukube et al. proposed the concept of an electrolarynx with a pitch control function [14]. Watson et al. investigated the improvement of the electrolaryngeal speech understanding by the fundamental frequency variation [15]. To improve the speech quality of laryngectomized speakers of Mandarin, an electrolarynx with tonal control function by using the movement of a trackball was developed [16]. After that, Wan et al. developed and evaluated a wheel-controlled pitch-adjustable electrolarynx for laryngectomized Mandarin speakers [17].

A wearable electrolarynx was also investigated [18–20]. Goldstein proposed and validated a hands-free electrolarynx triggered by neck-muscle electromyographic activity [18]. Yabu examined adding tonation to the input signal of an electric artificial larynx controlled by finger pressure [20].

To address the unnaturalness of the electrolaryngeal speech and the annoyance caused by the sound source signals of the electrolarynx, Nakamura et al. proposed speaking-aid systems, in which the electrolaryngeal speech is converted to normal speech by the proposed voice conversion method. Supin et al. also suppressed the radiated noise from the electrolarynx by the multiband time-domain amplitude modulation. Fuchs et al. presented a method to automatically learn an artificial F_0-contour to improve artificial larynx transducer speech [23].

As described above, various investigations have been performed in order to improve the quality of electrolarynx speech and usability. However, electric artificial laryngeal speech is still in the development phase, and artificial speech is difficult to clearly transmit to another person. When using an electrical artificial larynx, in addition to the uttered speech, the vibration sound of the vibrator is also heard. Therefore, increasing the output of the oscillator does not necessarily improve the sound quality. In addition, the maximum output level of the transducer is limited. For these reasons, in order to improve the sound quality, it is necessary to effectively improve the frequency characteristics of the input signal.

In the present study, first, by comparing the output frequency characteristics of sounds that are uttered naturally and using an electrical artificial larynx, an optimizing filter that enables the artificial voice to be heard more naturally and clearly is designed. Sounds uttered using filtered and non-filtered input signals are compared to naturally spoken voices, and the effectiveness of the filter is evaluated. In addition, a subjective evaluation is carried out on the filtered speech.

2 Filter Design and Validation

2.1 Basic Source-filter Theory

The source-filter model was first proposed by Chiba et al. [24], who explained the theory of speech production clearly by modeling an acoustic system using the excitation of vibration at the vocal cords, sound transmission via the vocal tract, and sound radiation from the mouth, as a simple linear system. In the present study, we consider an optimal voice production system using an electric artificial larynx based on the source-filter theory. A schematic diagram of the source-filter model is shown in Fig. 1. Here, s in the figure indicates the frequency. In the source-filter model, the process of speech production can be modeled as a composition of the following characteristics: the frequency characteristics $U(s)$ of the sound excited at the vocal cords, a linear time-invariant transmission system $T(s)$ via the vocal tract, and the sound radiation characteristics $R(s)$ for the mouth. Therefore, the spectrum of an uttered vowel, $P(s)$, is expressed as follows:

$$P(s) = U(s)T(s)R(s) \tag{1}$$

As is clear from the above mentioned modeling method, it is possible to separate the frequency characteristics of the sound source excited at the vocal cords $U(s)$ and the other transfer characteristics of $T(s)R(s)$. Considering that most laryngectomy patients have vocal tracts that are in good condition, the difference between the real voice and the artificial voice produced by the electric artificial larynx is considered to be only the excitation mechanism for the sound source. Consequently, based on the source-filter theory, if the sound source produced near the vocal cords by the electric artificial larynx is tentatively the same as the originally excited sound of the real voice, the uttered voice theoretically has the same frequency characteristics as the real voice. However, in most cases, artificial voices produced by an electric artificial larynx, which should have a sound source simulating that of real vocal cords, are difficult to understand.

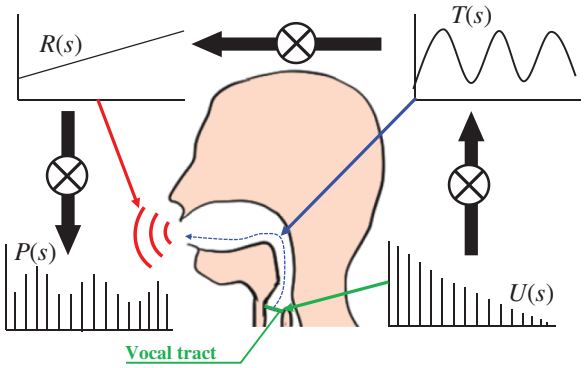

Figure 1: Schematic diagram of the source-filter model with the frequency characteristics of each acoustic element: the sound excited at the vocal cords $U(s)$, the linear time-invariant transmission system $T(s)$ via the vocal tract, and the sound radiation characteristics $R(s)$ from the mouth

In the present study, in order to investigate the reason for the unclearness of an artificial voice, the vocal-tract characteristics for each vowel uttered by placing a vibration speaker in contact with the surface of the neck and those uttered by a real voice were first compared. Then, considering the difference between the two envelopes of the frequency characteristics of artificial and real voices, a filter, which enables the artificial voice to be heard more clearly, was designed, and the improvement in the sound quality of the artificial voice with the designed filter was investigated.

2.2 Measurement of Speech Characteristics

In order to compare the envelopes of the frequency characteristics, real and artificial voices uttering the five Japanese vowels /a/, /i/, /u/, /e/, and /o/ were first recorded.

The excitation for the artificial voice was performed by inputting a positive ramp sawtooth wave (fundamental frequency: 120 Hz, sampling frequency: 48,000 Hz, 16 bit) into a vibration speaker (Fig. 3a, Hanwha, UMA-BVS03). The output level from the speaker was determined so that the voice uttered by the excitation of the speaker should be as loud as possible. Then, the input level into the speaker was carefully adjusted so as to avoid excessive excitation. Note that the input level into the speaker was fixed in all of the investigations of the present paper. The waveform and frequency characteristics of the adopted excitation source are shown in Figs. 2a and 2b, respectively. The waveform of each vowel, which was uttered for 10 seconds, was recorded using a directional microphone (Fig. 3b, AKG, C414 XLII). The distance between the speaker and the directional microphone was set to one meter, assuming ordinary face-to-face conversation. On the other hand, the real voice was uttered so that the equivalent sound pressure level averaged for ten seconds should be approximately 70 dB at a point one meter from the mouth. Using the above mentioned measurement scheme, real and artificial voices of one male in his twenties were recorded and were used for the investigation.

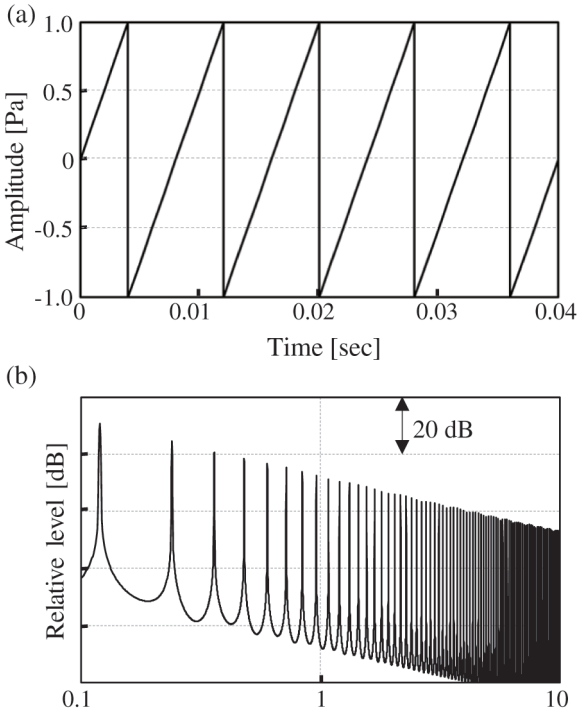

Figure 2: (a) Waveform and (b) frequency characteristics of the positive ramp sawtooth wave adopted to excite the artificial voice

2.3 Comparison of Natural and Artificial Voices

First, the vocal-tract characteristics of voices were evaluated using the spectral envelopes extracted by linear predictive coding (LPC). An example of the correspondence between the original frequency characteristics and their LPC envelope, when /a/ is uttered by a real voice, is shown in Fig. 4 as a red line. As indicated in the figure, the two peaks of the LPC envelope correspond to the resonance frequencies of the vocal tract, i.e., the first formant (F1) and the second formant (F2), as shown by the

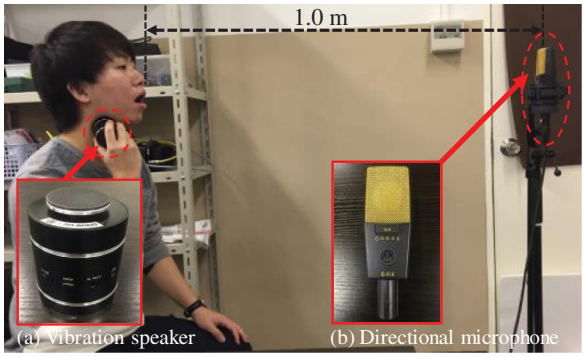

Figure 3: Measurement situation of the artificial voice involving exciting the neck by (a) a vibration speaker. The uttered voice was recorded using (b) a directional microphone placed at a receiving point one meter from the mouth

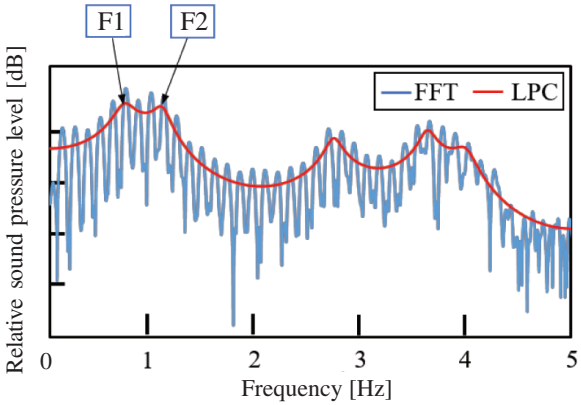

Figure 4: Correspondence between the original frequency characteristics of the natural voice for /a/ and the linear predictive coding (LPC) envelope

arrows in the figure. Second, the measurement results for the real and artificial voices for five vowels are shown in Fig. 5. Comparing the LPC envelopes of the real and artificial voices for /a/ in Fig. 5a, the real voice includes relatively broad frequency components in the range from 500 Hz to 1 kHz, whereas the frequency characteristics of the artificial voice around that frequency range are discrete and sharp. As a result, the peak levels for the first and second formants of the artificial voice are relatively low compared to those for the real voice. As is also clear from the results for other vowels, shown in Figs. 5b–5e, the peak levels for the first and second formants for artificial voice indicate relatively narrow-band characteristics and therefore had lower values compared to the real voice. Since the frequency components of F1 and F2 have an important role in identifying uttered vowels, the artificial voice, which generally has a lower sound pressure level in the frequency range around F1 and/or F2, may have a lower sound pressure level than the real voice. Therefore, in order to enhance the frequency components around F1 and F2, the following three filters were designed. The frequency characteristics of the designed filters are shown in Fig. 6. Hirahara et al. reported the mean frequency of F1 and F2 for the five Japanese vowels [25] for an adult male and an adult female, as indicated in Fig. 7. The formant frequencies of F1 and F2 and those shown in Fig. 7 for the adult male are similar. According to Fig. 7, F1 and F2 for a vowel uttered by an adult male is in the frequency range from 200 Hz to approximately 2.2 kHz. Filter 1, which covers the entire frequency range from 200 to 2.2 kHz, was initially set as shown in Fig. 6a. Note that a male is the target gender in the present paper because, in a basic study, it is difficult to consider

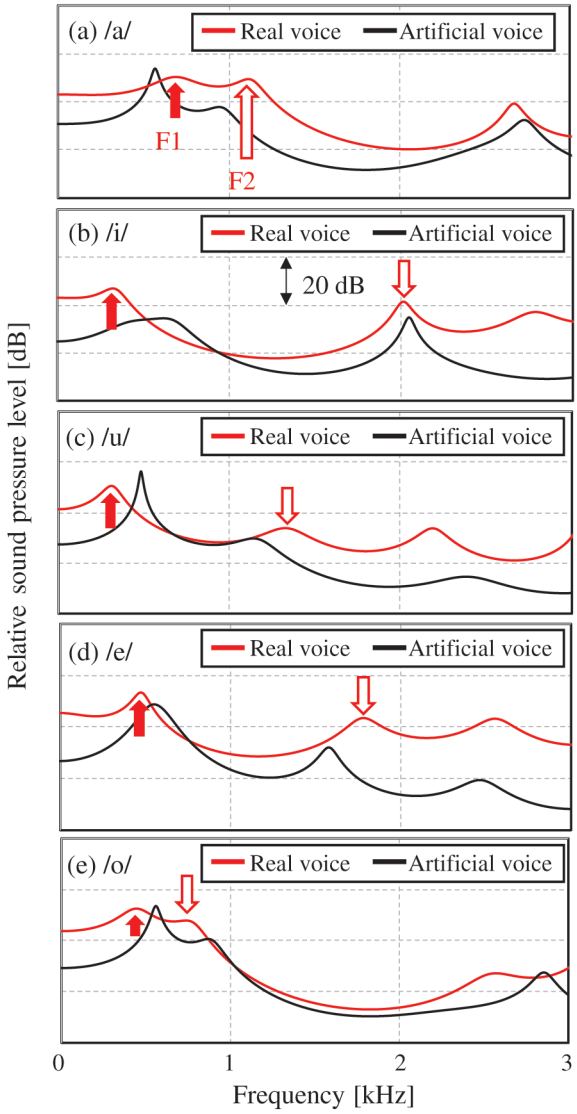

Figure 5: Measurement results for the LPC envelopes of the real and artificial voices for the five Japanese vowels: (a) /a/, (b) /i/, (c) /u/, (d) /e/, and (e) /o/

such a broad frequency range that covers male and female voices. However, the frequency range covered by Filter 1 is too broad, and it is necessary to verify the effect of enhancing the frequency components more efficiently. For that reason, by dividing the five vowels into two groups of /a/ and /o/ and /i/, /u/, and /e/, two filters, i.e., Filters 2 and 3, which cover the former two vowels and the latter three vowels, were set, as shown in Figs. 6b and 6c, respectively. Filter 2 covers the frequency range from 400 Hz to 1.2 kHz for /a/ and /o/, whereas Filter 3 covers that from 200 Hz to 600 Hz and from 1.2 kHz to 2.2 kHz for /i/, /u/, and /e/. The details of the filtering process using these three filters are described in the next section.

2.4 Effect of Optimal Filters on the Artificial Voice

The input signal of a sawtooth wave was filtered through Filters 1, 2, and 3 generated in the previous section using Adobe Audition. However, at this point, only the components of the frequency range specific to each filter were filtered, because the maximum level for all three filters is 0 dB, as shown in

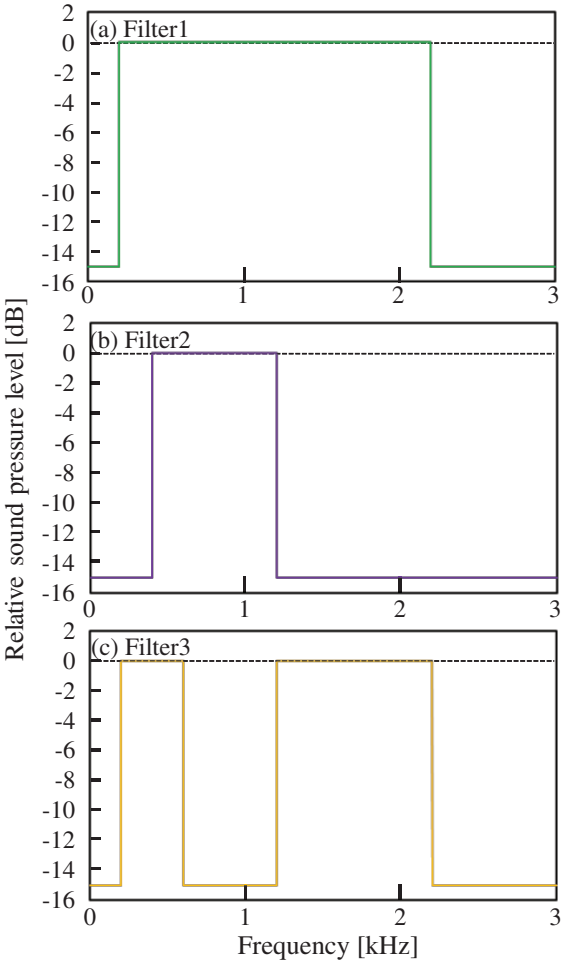

Figure 6: Frequency characteristics of the designed filters: (a) Filter 1, (b) Filter 2, and (c) Filter 3

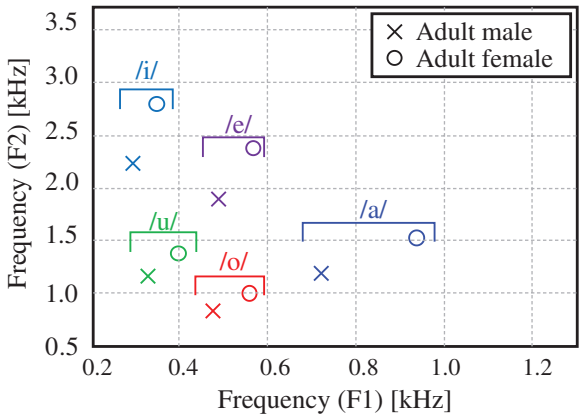

Figure 7: Frequencies of the first (F1) and second (F2) formants of the five Japanese vowels for an adult male and an adult female [15]

Fig. 6. If these filtered voices were to be directly evaluated, the filtered results obtained using Filter 1, including the broadest frequency components, should be evaluated as the best. However, such considerations produce only predictable results. Therefore, the amplitudes of these waveforms *Sig*(*t*) were adjusted so that their sound energy levels L_E, which were calculated by summation of *Sig*(*t*) for ten seconds (480,000 samples) as follows:

$$L_E = 10 \log \sum_{i=1}^{480000} (sig(i\Delta t))^2 \tag{2}$$

were equal among respective filter conditions. The reason for performing the above mentioned signal operation is that, when a stationary excitation is actually performed by a speaker, the maximum energy level of the stationary signal that can be input to the speaker is constant in each device. Then, the total sound energy level of the reproduced signals should be equally fixed to the same value in all filter conditions, because all the filtered signals should be equally evaluated. The frequency characteristics of the original sawtooth wave and the three input signals generated by the above mentioned treatment are shown in Fig. 8. For ease of understanding, only the envelope of each frequency characteristic is shown here. The signal filtered by Filter 2 naturally has a maximum level in the entire frequency range, because the enhanced frequency range is the narrowest among the three filters.

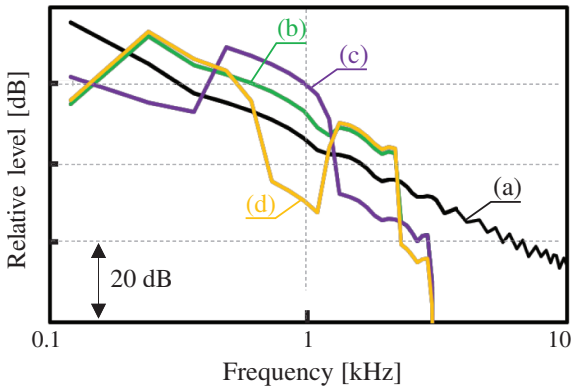

Figure 8: Frequency characteristics of (a) the original sawtooth wave and three input signals filtered by (b) Filter 1, (c) Filter 2, and (d) Filter 3

The voices uttered using the four input signals consisting of the original sawtooth waveform and the other three signals filtered through Filters 1, 2, and 3 were recorded. The LPC envelopes of these four signals are shown for each of the five vowels in Figs. 9a through 9e. In Figs. 9a and 9e, the sound pressure levels around F1 and F2 using Filter 2 (purple line) are enhanced compared to those of the original artificial voice (black line). In contrast, in Figs. 9b–9d, the sound pressure levels around F1 and F2 using Filter 3 (yellow line) are enhanced.

3 Subjective Evaluation Test

Subjective evaluation experiments were conducted to evaluate the intelligibility, naturalness, and loudness of the voices of which the formant components were enhanced by filtering. The details and results are described hereinafter.

3.1 Procedure

A subjective evaluation experiment was conducted on the voices obtained using the generated input signals. The subjects of the experiment were 10 males in their twenties. This experiment consists of three

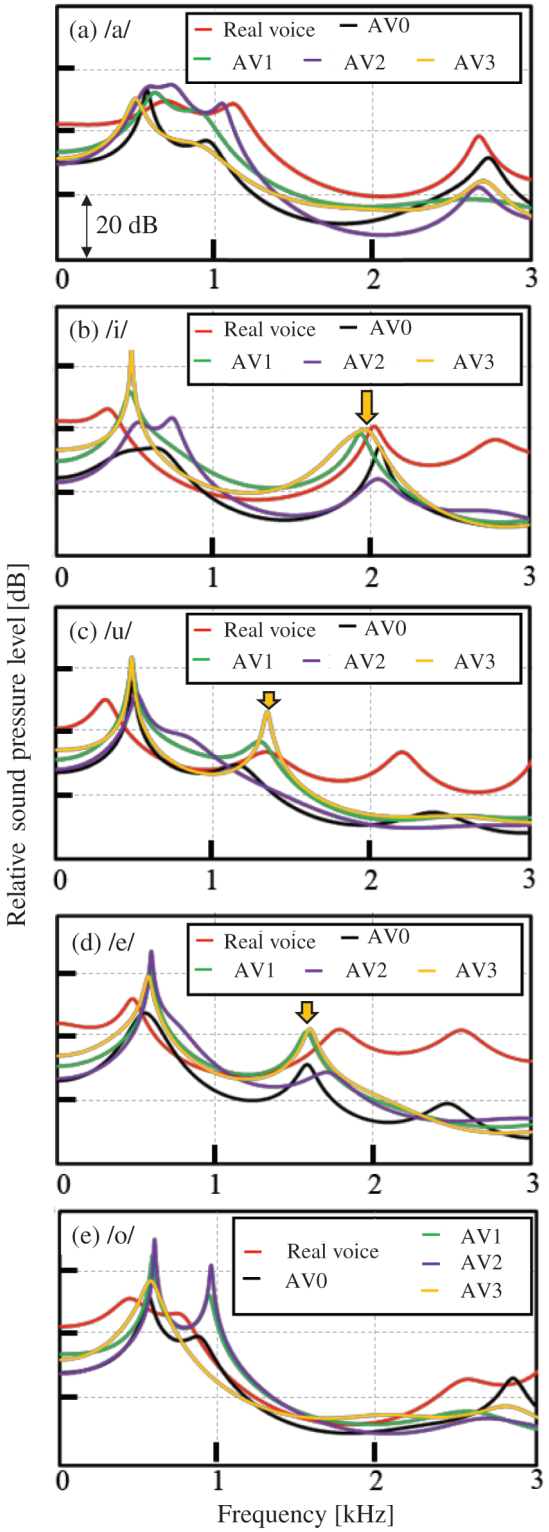

Figure 9: Measurement results for the real and artificial voices (AV0, AV1, AV2, and AV3) for the five Japanese vowels: (a) /a/, (b) /i/, (c) /u/, (d) /e/, and (e) /o/

parts. In Experiment 1, the recorded voices described in Section 2.4 were presented to the subjects through headphones. In Experiment 2, the subjects evaluated their artificial voices for the five Japanese vowels uttered using the vibration speaker. Note that all subjects were given sufficient time to practice before the main experiment so that they could appropriately produce the artificial voices. In Experiment 3, vowels uttered by other subjects in Experiment 2 were presented, which were evaluated by the subjects of Experiment 3. Then, in Experiments 2 and 3, the ten subjects were arranged in five pairs, who alternately conducted Experiments 2 and 3 for each pair. In Experiments 2 and 3, the distance between the speaker's mouth and the microphone in the recording and that between the speaker and the listener were set to one meter. In Experiments 1 and 3, the subjects were first presented with real voices as a reference, which were then compared to the artificial voices. Note that the real voices were controlled to be uttered by the subjects with sound pressure levels ranging from 69 dB to 77 dB at a point one meter from the mouth. After that, the artificial voice produced by the original sawtooth wave (AV0), that produced by the filtered sawtooth wave using Filter 1 (AV1), that produced by the filtered sawtooth wave using Filter 2 (AV2), and that produced by the filtered sawtooth wave using Filter 3 (AV3) were evaluated by the subjects. The intelligibility, naturalness, and loudness of the uttered voices were rated using seven monopolar categories. In each experiment, the subjects indicated their ratings of the intelligibility, naturalness, and loudness of the voices (AV1, AV2, and AV3) so that their ratings under the condition of AV0 should be "4."

3.2 Results and Discussion

Fig. 10 shows the arithmetically averaged values and their standard deviations for the evaluated values of (1) intelligibility, (2) naturalness, and (3) loudness in Experiments 1, 2, and 3, respectively. In addition, the significance of the differences between each condition was calculated by using the Wilcoxon signed-rank test, and the conditions with $p < 0.05$ are indicated with * mark in the figure. Noted that the method was adopted because all of the scores of the conditions without the filters were fixed to "4," and they have a variance of zero. The results of each experiment show that the rating for intelligibility and naturalness for /a/ and /o/ increased under condition AV2 compared to condition AV0. They also show the significant differences as shown in the figures of Ex.1-e, Ex.2-a, Ex.2-e, Ex.3-a, Ex.3-e. This indicates not only that the utterance of another subject, but also the utterance of oneself, became difficult to clearly understand. In contrast, the evaluated ratings for /i/, /u/, and /e/ were not particularly high compared to those for condition AV0. This is because the input signal of AV2 is optimized to the utterances of /a/ and /o/. Only in the condition of Experiment 2, the rating of the loudness for condition AV2 is lower than those of the intelligibility or naturalness. The reason may be that, in the case of hearing one's own voice, the influence of the sound that directly propagates from the vibration speaker is greatly emphasized compared to that of the voice uttered from the mouth. However, even if an increase in loudness was observed, it is appropriate that only the intelligibility and naturalness of the artificial voice can be increased from the viewpoint of the ease of face-to-face conversation. In contrast, the averaged ratings of all results for /i/, /u/ and /e/ were lower than for the other filter conditions. The enhanced frequency range may be considered not to match the frequencies of F1 and F2. On the other hand, with regard to the conditions of /i/, /u/, and /e/, no particular improvement in intelligibility was observed for condition AV3. Because the Filters 1 and 3 have a relatively wider frequency range compared to the Filter 2, the level of the amplitude of the filtered excitation waveforms by the Filters 1 and 3 are relatively lower than that filtered by the Filter 2 considering the evaluation method based on Eq. (2). Then, the level difference for the uttered voice between AV0 and AV3 resulted in a slight increase, which did not improve the evaluated score. However, in some cases for naturalness in Experiment 2, the ratings for /i/, /u/, and /e/ were slightly higher, which may be due to the increase of the output sound pressure levels around F2 for AV3, as indicated by the arrows in Fig. 9.

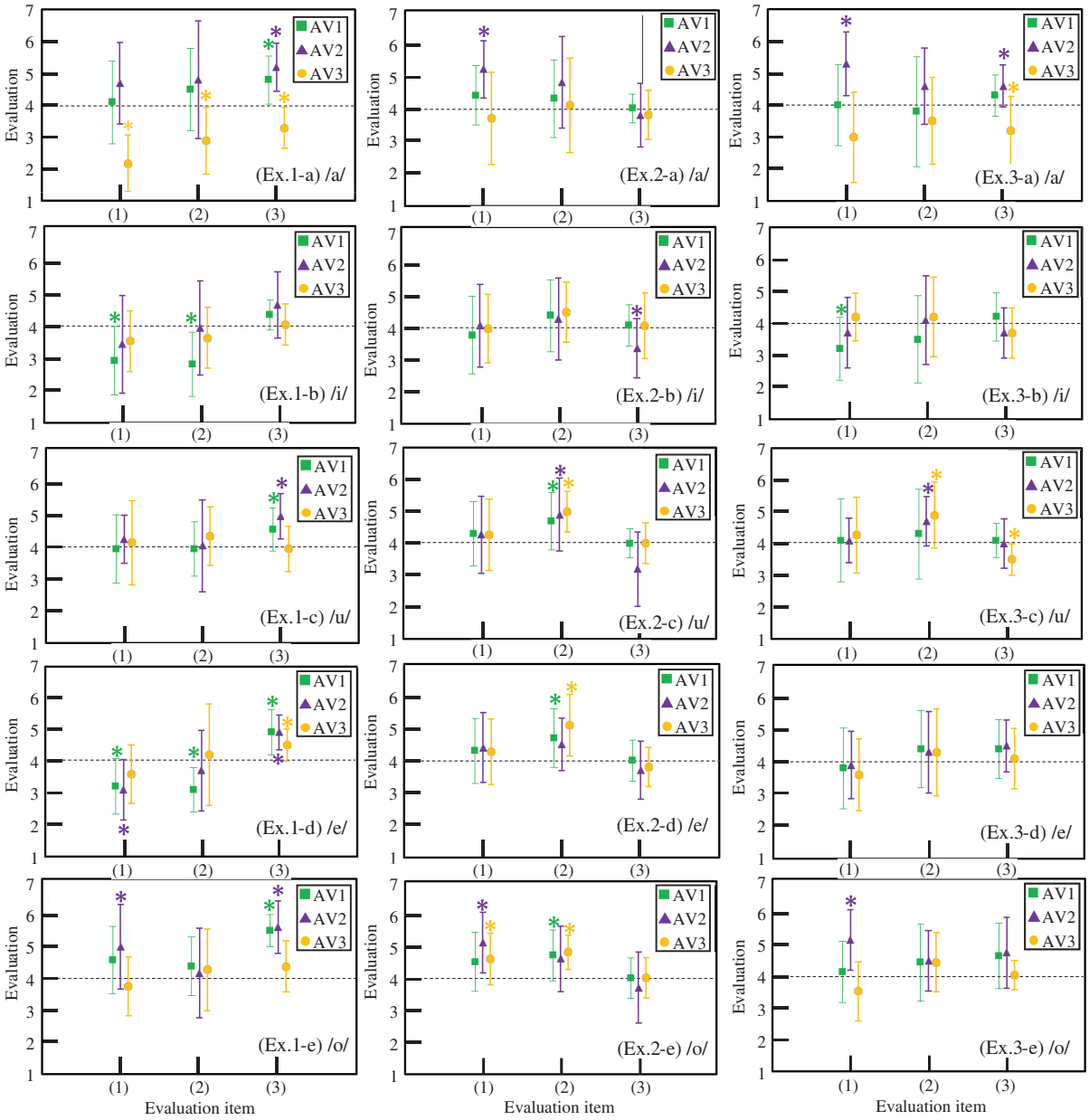

Figure 10: Results of the subjective evaluation test including the mean values and their standard deviations for the evaluated ratings of the three evaluation items (* $p < 0.05$) of (1) intelligibility, (2) naturalness, and (3) loudness in Experiment 1 from (Ex. 1-a) through (Ex. 1-e), Experiment 2 from (Ex. 2-a) through (Ex. 2-e), and Experiment 3 from (Ex. 3-a) through (Ex. 3-e)

Next, the physical characteristics of the uttered voices were evaluated, and compared to the above mentioned subjective results. As the physical parameters, which can be related to the loudness and intelligibility of the uttered voices, following two kinds of parameters A and B were obtained.

A. The RMSs of the waveforms of the uttered voices by each of the signals of AV1, 2 and 3 were calculated, and the difference of the RMSs between the conditions of AV0, and AV1/2/3 were obtained as ΔL_{RMS}.

B. The correlation coefficients between the LPC of the uttered waveform by each of the signals of AV0, 1, 2 and 3 and the real voices were obtained, and the difference of the correlation coefficients between the conditions of AV0, and AV1/2/3 were obtained as $\Delta Corr$.

Figs. 11a and 11b indicate the calculated ΔL_{RMS} and $\Delta Corr$, respectively. Firstly, as can be seen in the results of "loudness" of Ex.1-a through Ex.1-e of Fig. 10, the evaluated results of the loudness of AV2 indicate the highest score, and those of AV3 indicate the lowest scores. These tendencies agree with the ΔL_{RMS} of Fig. 11a. Next, as can be seen in Fig. 11b, the $\Delta Corr$s of the vowels of /i/, /u/, and /e/ indicate relatively higher scores in the condition of AV3, whereas the $\Delta Corr$s of the vowels of /a/, and /o/ indicate relatively higher scores in the condition of AV2. These tendencies also agree with the above mentioned characteristics of the intelligibility for the cases of /a/ and /o/, whereas the intelligibility of the uttered vowels of /i/, /u/, and /e/ were not sufficiently improved in the subjective results.

Therefore, we conclude that the quality of the utterances of /a/ and /o/ generally increased. However, the quality of the utterances of /i/, /u/, and /e/ did not clearly increase, and some issues remain regarding the integrated optimization of the filtering scheme for all of the vowels.

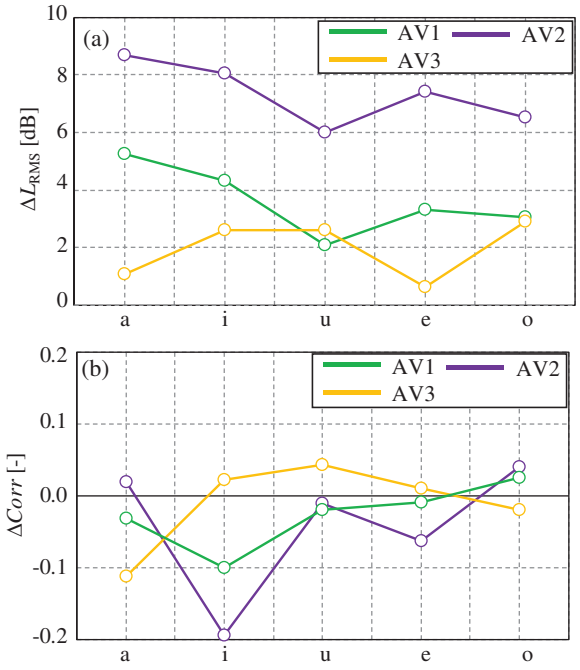

Figure 11: Results of the physical characteristics of the uttered voices in the conditions of AV1, 2, and 3. (a) ΔL_{RMS}, and (b) $\Delta Corr$

4 Conclusions

In order to improve the sound quality of vocalization using an electrical artificial larynx, the effect of optimizing the input signal was investigated. Based on a comparison of the frequency characteristics of the real and artificial voices, an optimal filter that can make the frequency characteristics of the artificial voice closer to the real voice was generated, and the influence of this filter on the improvement of the quality of the artificial voice was investigated. First, as a basic study, by comparison of five Japanese vowels artificially uttered using a vibrating speaker and by a real voice, the sound pressure levels of the artificial voices at the first and second formants, which greatly influence the intelligibility of vowels, were lower than those of a natural voice. Therefore, three filters were generated, each of which enhanced (1)

all five vowels, (2) /a/ and /o/, or (3) /i/, /u/, and /e/. In order to evaluate the artificial voice when the signals were filtered, measurement of the physical vocal-tract characteristics and a subjective evaluation experiment were carried out. The results indicated that the sound pressure levels for the first and second formants of the artificial voice were improved for all five vowels. Based on the subjective evaluation experiment, the intelligibility of the artificial voices of /a/ and /o/ was improved, whereas little improvement was observed in the case of the vowels /i/, /u/, and /e/. Based on these results, the effect of changing the input signal into the vibration speaker depending on the vowels was confirmed, whereas not all conditions were improved. In the future, we intend to develop an efficient boosting system that can smoothly and automatically convert the individual speech signals in a usual conversation, in which vowels are randomly uttered.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Barney, H. L., Haworth, F. E., Dunn, H. K. (1959). An experimental transistorized artificial larynx. *Bell System Technical Journal, 38(6),* 1337–1356. DOI 10.1002/j.1538-7305.1959.tb01591.x.
2. Holley, S. C., Lerman, J., Randolph, K. (1983). A comparison of the intelligibility of esophageal, electrolaryngeal, and normal speech in quiet and in noise. *Journal of Communication Disorders, 16(2),* 143–155. DOI 10.1016/0021-9924(83)90045-X.
3. Weiss, M. S., Yeni-Komshian, G. H., Heinz, J. M. (1975). Acoustic characteristics of speech produced with an electronic artificial larynx. *Journal of the Acoustical Society of America, 58(S1),* S112–S112. DOI 10.1121/1.2001866.
4. Weiss, M. S., Yeni-Komshian, G. H., Heinz, J. M. (1979). Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. *Journal of the Acoustical Society of America, 65(5),* 1298–1308. DOI 10.1121/1.382697.
5. Liu, H., Zhao, Q., Wan, M., Wang, S. (2006). Application of spectral subtraction method on enhancement of electrolarynx speech. *Journal of the Acoustical Society of America, 120(1),* 398–406. DOI 10.1121/1.2203592.
6. Liu, H., Zhao, Q., Wan, M. (2006). Enhancement of electrolarynx speech based on auditory masking. *IEEE Transactions on Biomedical Engineering, 53(5),* 865–874. DOI 10.1109/TBME.2006.870236.
7. Liu, H., Zhao, Q., Wan, M., Wang, S. (2006). Application of spectral subtraction method on enhancement of electrolarynx speech. *Journal of the Acoustical Society of America, 120(1),* 398–406. DOI 10.1121/1.2203592.
8. Basha, S. K., Pandey, P. C. (2012). Real-time enhancement of electrolaryngeal speech by spectral subtraction. *Proceedings of the 18th National Conference on Communications,* 516–520.
9. Matsui, K., Otsuki, Y., Nakatoh, Y., Kato, Y. O. (2013). A preliminary user interface study of speech enhancement system. *Proceedings of the 1st International Conference on Industrial Application Engineering 2013*, pp. 53–56.
10. Tanaka, K., Toda, T., Neubig, G., Sakti, S., Nakamura, S. (2013). A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion. *Proceedings of Interspeech,* pp. 3067–3071.
11. Tanaka, K., Toda, T., Neubig, G., Sakti, S., Nakamura, S. (2014). An evaluation of excitation feature prediction in a hybrid approach to electrolaryngeal speech enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 4488–4492.
12. Meltzner, G. S., Kobler, J. B., Hillman, R. E. (2003). Measuring the neck frequency response function of laryngectomy patients: implications for the design of electrolarynx devices. *Journal of the Acoustical Society of America, 114(2),* 1035–1047. DOI 10.1121/1.1582440.

13. Wu, L., Xiao, K., Dong, J., Wang, S., Wan, M. (2014). Measurement of the sound transmission characteristics of normal neck tissue using a reflectionless uniform tube. *Journal of the Acoustical Society of America, 136(1),* 350–356. DOI 10.1121/1.4883355.
14. Ifukube, T., Uemi, N. (1999). A new electrical larynx with pitch control function. *Second East Asian Conference on Phonosurgery,* 7−10.
15. Watson, P. J., Schlauch, R. S. (2009). Fundamental frequency variation with an electrolarynx improves speech understanding: a case study. *American Journal of Speech-Language Pathology, 18(2),* 162–167.
16. Wan, C., Wang, E., Wu, L., Wang, S., Wan, M. (2012). Design and evaluation of an electrolarynx with tonal control function for Mandarin. *Folia Phoniatrica et Logopaedica, 64(6),* 290–296. DOI 10.1159/000346861.
17. Wang, L., Feng, Y., Yang, Z., Niu, H. (2017). Development and evaluation of wheel-controlled pitch-adjustable electrolarynx. *Medical & Biological Engineering & Computing, 55(8),* 1463–1472. DOI 10.1007/s11517-016-1606-6.
18. Goldstein, E. A., Heaton, J. T., Kobler, J. B., Stanley, G. B., Hillman, R. E. (2004). Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. *IEEE Transactions on Biomedical Engineering, 51(2),* 325–332. DOI 10.1109/TBME.2003.820373.
19. Hashiba, M., Sugai, Y., Izumi, T., Ino, S., Ifukube, T. (2007). Development of a wearable electro-larynx for laryngectomees and its evaluation. *Conference Proceedings—IEEE Engineering in Medicine and Biology Society,* 5267−5270.
20. Yabu, K., Ifukube, T. (2016). Design of the wearable electrolarynx equipped with the loudspeaker. *Transactions of the Virtual Reality Society of Japan, 21(2),* 295–301.
21. Nakamura, K., Toda, T., Saruwatari, H., Shikano, K. (2012). Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. *Speech Communication, 54(1),* 134–146. DOI 10.1016/j.specom.2011.07.007.
22. Xiao, K., Wang, S., Wan, M. X., Wu, L. (2018). Radiated noise suppression for electrolarynx speech based on multiband time-domain amplitude modulation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(9),* 1585–1593. DOI 10.1109/TASLP.2018.2834729.
23. Fuchs, A. K., Hagmüller, M. (2012). Learning an artificial F_0-contour for ALT speech. *Proceedings of Interspeech,* 9–13.
24. Chiba, T., Kajiyama, M. (1941). *The vowel: its nature and structure*. Japan: Tokyo-Kaiseikan Pub. Co., Ltd.
25. Hirahara, T., Akahane-Yamada, R. (2004). Acoustic characteristics of Japanese vowels. *Proceedings of 18th International Congress on Acoustics,* pp. 3287–3290.