INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Speech clarity improvement by vocal self-training using a hearing impairment simulator and its correlation with an auditory modulation index

Toshio Irino[1], *Soichi Higashiyama*[2], *and Hanako Yoshigi*[3]

Faculty of Systems Engineering, Wakayama University, Japan
[1]irino@wakayama-u.ac.jp, [2]higashiyama.soichi@g.wakayama-u.jp ,
[3]yoshigi.hanako@g.wakayama-u.ac.jp

Abstract

We performed two experiments to ascertain whether vocal self-training improves speech clarity, particularly when the feedback speech is degraded by a hearing impairment simulator. Speech sounds before and after the training were recorded under noisy and quiet conditions and their speech clarity was evaluated by subjective listening tests using Scheffe's paired comparison. We also analyzed the auditory modulation features to derive an index to explain the subjective clarity scores. The auditory modulation index highly correlated with subjective scores and seems a good candidate for predicting speech clarity.

Index Terms: clear speech, gammachirp auditory filterbank, modulation filterbank,

1. Introduction

As the super-aging society approaches in many countries, it is essential to develop new-generation assistive devices and to understand how to speak to elderly persons with hearing loss (HL) even when they do not wear a device. While many previous studies have investigated clear speech, they have mainly been evaluated by normal hearing (NH) listeners e.g.,[1, 2, 3, 4, 5]. It has been reported that the modulation component in the envelope domain is an important feature for speech clarity and Lombard speech [6, 7]. However, there seems to be little knowledge regarding what objective measure can be reliably used to evaluate speech clarity for elderly persons.

In this study, we first performed vocal self-training experiments using a hearing impairment simulator (e.g., WHIS [8, 9, 10], see section 2.3) as shown in Fig. 1. The introduction of WHIS was intended to give the NH trainees experience of what elderly persons may hear and to lead them to improve their utterances to overcome the hearing difficulty. We then performed subjective evaluation and analysis of speech sounds recorded before and after the training with the aim of developing a reasonably objective index that highly correlates with subjective evaluation. Such an index may provide useful information for developing future assistive devices and for establishing vocal training methods.

2. Vocal self-training experiments

Experiments on vocal self-training were conducted to evaluate the effects on improving speech clarity of auditory feedback and training conditions. Speech sounds were recorded before and after the training for evaluation by humans (Sec 3) and an auditory model (Sec 4.1).

2.1. Training and recording setup

Figure 1 shows the experimental setup. The participant sat alone in front of a head and torso simulator (HATS, distance = 1.5 m) in a soundproof room. The participant was simply instructed to practice pronunciation toward the HATS, which imitated an elderly person with HL. There was no vocal training

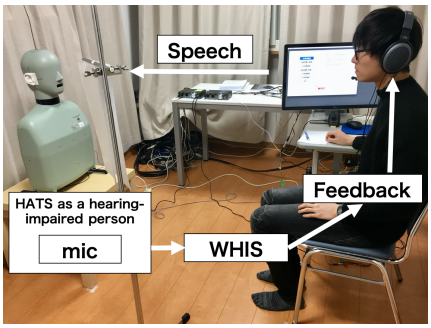


Figure 1: *Vocal self-training using Wadai Hearing Impairment Simulator (WHIS). See text for detail.*

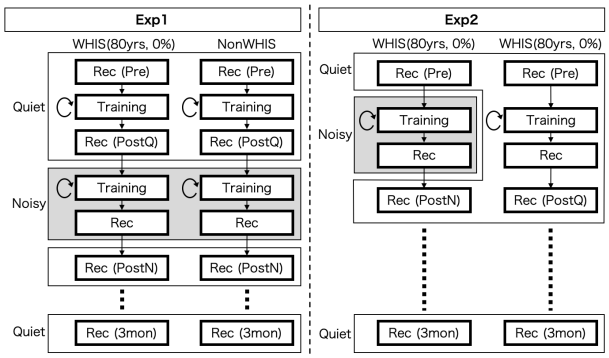


Figure 2: *Block diagrams of Exp. 1 and Exp. 2. "Rec" indicates the recording of a set of 20 words. "Train" indicates a training session using three words. The gray area indicates the "noisy" condition in which babble noise was played in the room. Otherwise, no external noise was played, i.e., "Quiet."*

instructor or manual for the training method. The participantfs speech sounds were recorded by a microphone located on the left auricle of the HATS. The sound was processed by WHIS [8, 9, 10] (see sec. 2.3) to simulate the average hearing level of an 80-year-old [11] and 0% compression healthiness. The processed sound was fed back to the participant to understand what elderly persons may hear. Before the training began, the participants recorded a set of words. They then proceeded to the training session to practice their own voice until the feedback speech was judged satisfactory. After that, they recorded the same word set again. The outputs of microphones placed outside and inside the HATS and near the speaker's mouth were recorded synchronously.

2.2. Experimental design

Figure 2 shows block diagrams of the two experiments. Experiment 1 was conducted to evaluate the effect of hearing impairment simulation on speech clarity improvement. Twenty

http://dx.doi.org/10.21437/Interspeech.2020-1081

participants were divided into two groups according to whether the feedback sounds were processed and degraded by WHIS or not (nonWHIS). Initially, the participant recorded a set of 20 words (Pre) from the list described in section 2.4, performed self-training using the first three words in the "Quiet" condition, and recorded the same set of 20 words again (PostQ). They then proceeded with self-training and recording under the "Noisy" condition with speech babble played from a loudspeaker in the room as background noise. The noise level $L_{eq} = 65$ dB at the output of HATS. Finally, they recorded the same set of 20 words (PostN) under the "Quiet" condition. Moreover, speech recording was performed after approximately three months (3Mon) to analyze the sustainability of the training effect.

Experiment 2 was conducted to separate the effect of background noise during the training. Twenty participants were divided into two groups. One group performed the training under the "Noisy" condition once; the other performed the training under the "Quiet" condition. In both cases, the feedback sounds were processed by WHIS. Unlike Exp.1, the training was not repeated.

2.3. Wadai Hearing Impairment Simulator, WHIS

Wadai (or Wakayama University) Hearing Impairment Simulator (WHIS) was developed to give NH listeners experience of what elderly persons may hear [8, 9, 10]. It is based on auditory processing using the compressive gammachirp auditory filter[12, 13, 14]. Briefly, the hearing level in the audiogram of a listener with hearing loss (HL) is decomposed into loss in OHC (outer hair cell, which enhances basilar-membrane vibration) [15] and loss in IHC (inner hair cell, which transduces vibration to neural firing) on a dB scale. Namely, $HL_{total} = HL_{OHC} + HL_{IHC}$ [16]. The OHC loss causes degradation of nonlinear compression applied by the auditory filter, while the IHC loss causes degradation of the filter gain. WHIS synthesizes simulated HL sound in accordance with the user setting. WHIS does not produce noticeable side-effect noise or distortion. Therefore, sound quality is maintained at a sufficiently high level to be usable in psychophysical experiments including this study. See [14] for more detail.

2.4. Word list and participants

A list of 20 words was presented to the participants. The list consisted of four-mora words drawn from a Japanese word database, FW03 [17], which contains lists controlled by familiarity. The familiarity level of the words used in this study was the second lowest. Note that one "mora" corresponds to approximately one CV syllable.

Twenty Japanese students participated in each training experiment after providing informed consent (40 in total; 20 males and 20 females between 21 and 23 years of age). They all had normal hearing thresholds of less than 20 dB HL between 125 and 8000 Hz. This experiment was approved by the local ethics committee of Wakayama University.

3. Subjective evaluation of speech clarity

Listening experiments were conducted to evaluate the training effect on speech clarity. Figure 3 shows the method.

3.1. Procedure

Speech sounds recorded at the microphone close to the mouth in the "Rec" stages in Fig. 2 were used for the evaluation. Speech sounds of six words were used for evaluation. They were selected from both trained words (1st, 2nd, and 3rd) and untrained words (4th, 12th, and 20th). The sound level, L_{eq}, was normalized to 70 dB to discourage the listeners from judgment based

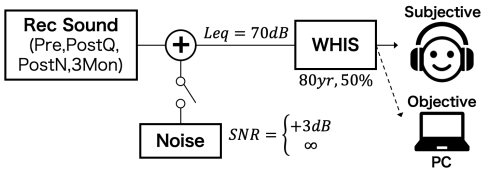
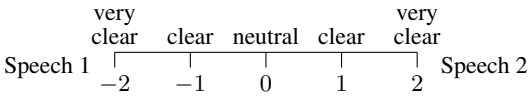


Figure 3: *Subjective and objective evaluation of speech clarity*

on the level because the post-trained sounds were almost always louder. The sound was mixed with babble noise with an SNR of +3dB. The no noise (SNR of ∞) condition was also included in the evaluation of Exp. 2 in Fig. 2. The mixed sound was then processed with WHIS with 80-year-old HL [11] and 50% compression healthiness to simulate elderly persons with HL. The listeners sat in a soundproof room and participated in an experiment using Scheffe's paired comparison [18]. They listened to pairs of speech sounds over headphones and judged the clarity on a 5-point scale ranging between -2 and +2, as shown below through a GUI.



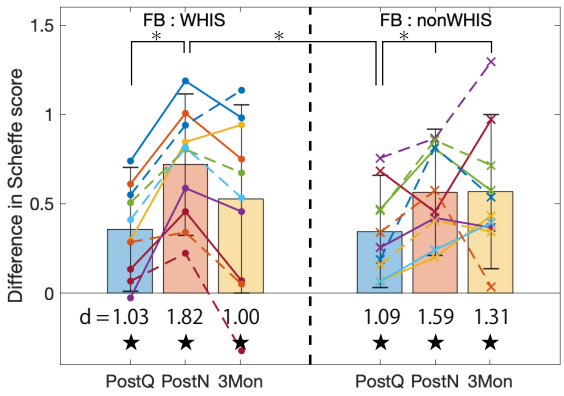
3.2. Condition and listeners

For the evaluation of Exp. 1 in Fig. 2, the total number of speech pairs was 1440, i.e., (6 words) $\times$ ($_4$C$_2$ combinations of training stages) $\times$ (2 permutation) $\times$ (20 speakers). For the evaluation of Exp. 2, the total number of speech pairs was also 1440, i.e., (6 words) $\times$ ($_3$C$_2$ combinations) $\times$ (2 permutation) $\times$ (20 speakers) $\times$ (2 listening SNRs). Ten Japanese students participated in each evaluation experiment after providing informed consent (20 in total; 10 males and 10 females between 20 and 23 years of age). They had not participated in the vocal self-training. They all had normal hearing thresholds of less than 20 dB HL between 125 and 8000 Hz. This experiment was approved by the local ethics committee of Wakayama University.

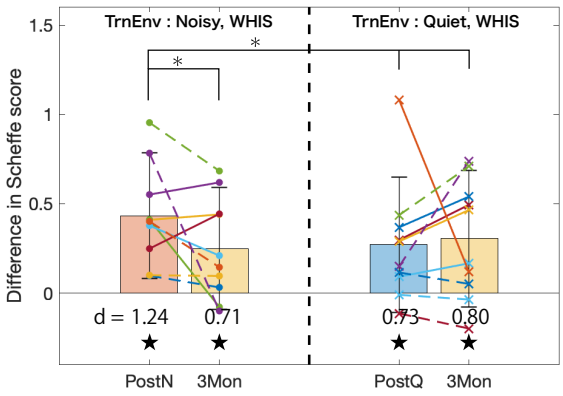
3.3. Results

Figure 4(a) shows the evaluation results for Exp. 1. The bar graph shows the difference in the average Scheffe scores [18] of 10 listeners between the pre-training (Pre) and post-training (PostN, PostQ, and 3Mon). The left three bars represent the differences when the feedback speech sounds were processed by WHIS, while the right three bars represent the differences when the sounds were not processed by WHIS (nonWHIS). The values were significantly different from zero for all cases (t-test,$\alpha = 0.05$). Cohen's d values [19] were always greater than 1.0 and the effect sizes are large. The value for PostN with WHIS was the largest (1.82). Multiple comparison with the Tukey-Kramer HSD method ($\alpha = 0.05$) [22] between the six conditions was performed and the asterisk (*) indicates a significant difference. The values in the PostN conditions are significantly different from those in the PostQ conditions for both WHIS and nonWHIS.

Figure 4(b) shows the evaluation results for Exp. 2. The average score changes were also significantly different from zero for all cases. Cohen's d value for PostN was 1.24, which means a large effect size. In contrast, the d values of the other conditions were between 0.70 and 0.80 (middle to large effect size) and less than the d values of Exp. 1 shown in Fig. 4(a). Multiple comparison with the Tukey-Kramer HSD method between the four conditions shows significant differences between PostQ and the other conditions. This implies that the training with



(a) *Evaluation results of Exp.1. Left: training with WHIS. Right: training without WHIS (nonWHIS).*



(b) *Evaluation results of Exp.2. Left: training under noisy condition with WHIS. Right: training under quiet condition with WHIS.*

Figure 4: *Evaluation results on speech clarity. The bar graph represents the difference in average Scheffe scores between speech sounds of pre-training (Pre) and post-training (PostN, PostQ, and 3Mon). The lines across training conditions show the score changes for individual listeners. The star ($\star$) shows significant difference from zero (t-test, $\alpha = 0.05$). Cohen's d value [19] is also presented above it. Multiple comparison with the Tukey-Kramer HSD method ($\alpha = 0.05$) between the six (or four) conditions was performed and the asterisk (*) indicates a significant difference.*

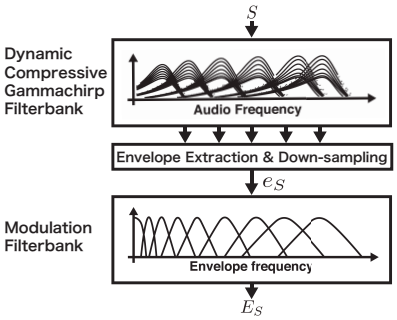


Figure 5: *Block diagram of envelope analysis using the dynamic compressive gammachirp filterbank (dcGC-FB) and modulation filterbank (MFB). See [20, 21] for more detail.*

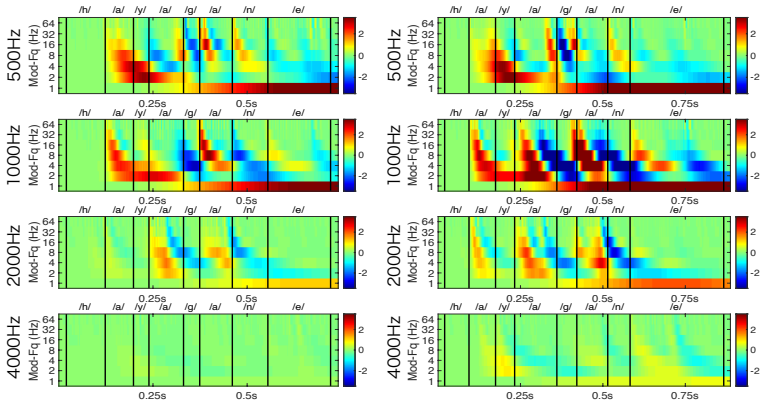


Figure 6: *Example of the modulation filterbank output of Fig. 5 when analyzing a word /hayagane/. Left panels show the results of the pre-training (Pre) speech. Right panels show the results of post-training (PostN) speech of Exp. 1. Abscissa: time. Ordinate: modulation frequency. The phonetic boundaries are shown by the vertical black lines. From upper panel to lower panel, the center frequencies of the auditory filters in the dcGC-FB are 505 Hz, 1023 Hz, 2039 Hz, and 4004 Hz.*

noisy background was effective.

In summary, the results imply that vocal self-training is an effective way to improve speech clarity and that effect is sustainable. The repeated training as in Exp. 1 improve that effectiveness. The training with a noisy background may also induce Lombard effect[6].

4. Auditory index for speech clarity

It is important to develop an objective measure or index for speech clarity. Such an index would be useful not only for vocal training methods but also for developing speech enhancement algorithms. In this section, we demonstrate an index based on auditory features that has good correlation with the subjective evaluation results described in section 3.

4.1. Auditory feature extraction

There are several speech intelligibility indices based on auditory models[23, 24, 25, 20, 26, 21]. They basically consist of an auditory filterbank, envelope extraction, and modulation frequency analysis to extract features of speech sounds. In this study, the filterbank structure used in the Gammachirp Envelope Distortion Index (GEDI) [20, 26], which reasonably predicts the intelligibility of speech enhanced by recent noise reduction algorithms, was used to develop an index that correlates with speech clarity. Figure 5 shows the block diagram. The gammachirp auditory filterbank (dcGC-FB) [27]and a modulation filterbank (MFB) are cascaded to extract envelope modulation features. Figure 6 shows an example of the filterbank output of Fig. 5 when analyzing a word /hayagane/. There are differences in the modulation patterns before training (Pre, left panels) and after training (PostQ, right panels). The modulation depth is much deeper in PostQ. We assumed that the modulation depth may highly correlate with speech clarity.

4.2. Auditory modulation index

The same speech sounds used in the subjective evaluation (Fig. 3) were analyzed by the model in Fig. 5. The sounds were processed by WHIS. The dcGC-FB has 100 channels, $\{i|1 \leq i \leq 100\}$, equally spaced along the $\mathrm{ERB_N}$-number [28] and covers the speech range between 100 and 6,000 Hz. The temporal envelopes are calculated from the output of the individual auditory filter using a Hilbert transform and a low-pass filter with

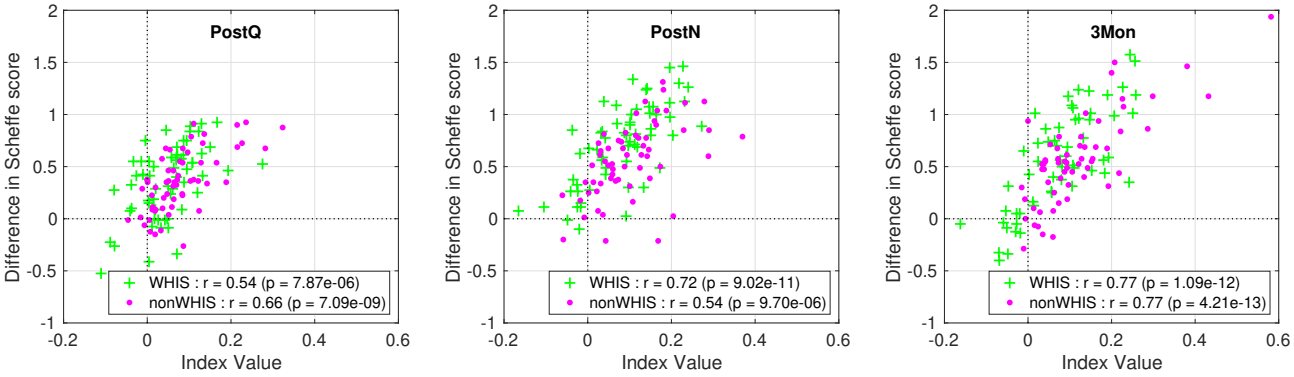


Figure 7: *Correlation between the index value and the difference in Scheffe score in Exp. 1 for PostQ, PostN, and 3Mon. Cross(+) shows the results under the WHIS condition; the filled circle shows the results under the nonWHIS condition. See Table 1 for a summary.*

a cutoff frequency of 150 Hz. The MFB consists of one low-pass filter (cutoff 1 Hz) and six octave bandpass filters (center frequencies between 2 Hz and 64 Hz), $\{j|1 \leq j \leq 7\}$. We calculated the rms value of the modulation filterbank output for individual mora in the four-mora words, $\{k|1 \leq k \leq 4\}$, to absorb the difference in speech duration. For this purpose, the phonetic boundary was determined by a segmentation method in Julius ASR [29] and compensated manually. The derived rms value is denoted as E_{ijk}. The values were calculated for speech sounds in Pre, PostQ, PostN, and 3Mon conditions. To evaluate the effect of training, we calculated the ratio between the Pre and Post conditions as

$$E_{ijk}^{(rat)} := \frac{E_{ijk}^{(Post)}}{E_{ijk}^{(Pre)}}. \quad (1)$$

Then, an auditory modulation index, I_{AM}, was defined as

$$I_{AM} = \log_{10}\left(\frac{\sum_{i=1}^{100}\sum_{j=2}^{7}\sum_{k=1}^{4} w_{ijk} \cdot E_{ijk}^{(rat)}}{100 \cdot 6 \cdot 4}\right) \quad (2)$$

where w_{ijk} is a weighting function and was set to unity ($w_{ijk} = 1$) in this study. The first channel of the MFB ($j = 1$) was excluded from the summation because the values were mainly a large DC bias in the envelope.

4.3. Correlation between subjective and objective results

We calculated correlation coefficients between the auditory envelope index, I_{AM}, in Eq. 2 and the subjective evaluation results described in section 3. Figure 7 shows scatter plots between the index values and the differences in the Scheffe scores for individual trainees' speech sounds in Exp. 1. All panels show positive correlations. The correlation coefficients were listed in the third row of Table 1. The values were greater than 0.5 and the maximum value was 0.77. The third row of Table 2 shows the correlation coefficients in Exp. 2. There were moderate to high correlations for both Exp. 1 and Exp. 2. The results imply that the auditory modulation index, I_{AM}, can be a good candidate for speech clarity evaluation.

4.3.1. Correlation with speech production parameters

We also calculated several speech parameter values, which have been believed to be correlated with speech clarity, from the speech sounds of Pre, PostN, PostQ, and 3Mon. The parameters were average speech speed per mora, average fundamental frequency ($\log_2 \bar{F}_o$), and open quadrant (O_q) of vocal fold. O_q can be estimated from the Electro-Glottal Graphh (EGG) waveform (see [30] for more detail). After plotting the graphs similar to Fig. 4, we found the average speech speed significantly decreased after the vocal training, the average fundamental frequency significantly increased, and O_q ratio significantly increased, in most cases [30]. Therefore, these parameters might reflect some aspects of speech clarity. To test the goodness, we also calculated correlation coefficients between these speech parameter values and the subjective evaluation results in section 3. The results are shown in the fourth row and below of Tables 1 and 2. The correlation coefficients were always less than those for I_{AM}. In many cases, the coefficients were less than 0.1, i.e., no correlation. There were only three conditions where the absolute value was greater than 0.4. Moreover, there is almost no consistency between the conditions. The indices based on these speech parameters were not sufficiently good for speech clarity evaluation although they are closely related to speech production.

Table 1: *Correlation coefficients in Exp. 1. Comparison between auditory index, speech speed, and F_o.*

Feedback	WHIS			NonWHIS		
Record	PostQ	PostN	3Mon	PostQ	PostN	3Mon
I_{AM}	**0.54**	**0.72**	**0.77**	**0.66**	**0.54**	**0.77**
Mora speed	-0.15	-0.14	-0.03	-0.48	-0.35	-0.37
$\log_2 F_o$	0.33	0.12	0.51	0.20	0.02	0.06

Table 2: *Correlation coefficients in Exp. 2. O_q were measured only in Exp.2. Feedback: WHIS.*

Training env.	Noisy		Quiet	
Record	PostN	3Mon	PostQ	3Mon
I_{AM}	**0.65**	**0.51**	**0.73**	**0.64**
Mora speed	0.03	-0.01	-0.31	-0.35
$\log_2 F_o$	-0.04	0.27	-0.09	0.48
$O_q$ratio	-0.10	0.02	0.02	0.02

5. Summary

In this study, we performed two experiments on voice self-training using WHIS and subjective evaluation of recorded speech sounds. Although the effect of the hearing impairment simulation was not very clear, repeated trainings under the noisy and quiet conditions were found to be effective. We also developed an auditory modulation index to predict speech clarity. The index was better than speech production parameters: speech speed, fundamental frequency, and open quadrant. Future work should improve the index by optimizing the weight coefficient, w_{ijk}, in Eq. 2.

Acknowledgements This research was supported by the JSPS KAKENHI Nos. JP16H01734 and JP18K10708, and NII CRIS collaborative research program operated by NII CRIS and LINE Corporation.

6. References

[1] J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 362–378, 2004.

[2] K. Maniwa, A. Jongman, and T. Wade, "Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1114–1125, 2008.

[3] ——, "Acoustic characteristics of clearly spoken english fricatives," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3962–3973, 2009.

[4] R. Smiljanić and A. R. Bradlow, "Speaking and hearing clearly: Talker and listener factors in speaking style changes," *Language and linguistics compass*, vol. 3, no. 1, pp. 236–264, 2009.

[5] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, 2014.

[6] H. Lane and B. Tranel, "The lombard sign and the role of hearing in speech," *Journal of Speech and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.

[7] H. R. Bosker and M. Cooke, "Talkers produce more pronounced amplitude modulations when speaking in noise," *The Journal of the Acoustical Society of America*, vol. 143, no. 2, pp. EL121–EL126, 2018.

[8] T. Irino, T. Fukawatase, M. Sakaguchi, R. Nisimura, H. Kawahara, and R. D. Patterson, "Accurate estimation of compression in simultaneous masking enables the simulation of hearing impairment for normal-hearing listeners," in *Basic Aspects of Hearing*. Springer, 2013, pp. 73–80.

[9] M. Nagae, T. Irino, R. Nisimura, H. Kawahara, and R. D. Patterson, "Hearing impairment simulator based on compressive gammachirp filter," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2014, pp. 1–4.

[10] "http://www.cs.tut.ac.jp/∼matsui/whis/ (last view on 26 jul 2020)."

[11] T. Tsuiki, S. Sasamori, N. Nankitsu, K. Ichihe, K. Murai, M. Murai, and K. Kawashima, "Study on age effect of hearing level of japanese," *Audiology Japan*, vol. 45, no. 3, pp. 241–250, 2002 (in Japanese).

[12] T. Irino and R. D. Patterson, "A time fomain, level-dependent auditory filter: the gammachirp," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 412–419, 1997.

[13] ——, "A compressive gammachirp auditory filter for both physiological and psychophysical data," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2008–2022, 2001.

[14] ——, "The gammachirp auditory filter and its application to speech perception," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 99–107, 2020.

[15] J. Pickles, *An introduction to the physiology of hearing*. Brill, 2013.

[16] B. C. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

[17] S. Amano, S. Sakamoto, T. Kondo, and Y. Suzuki, "Development of familiarity-controlled word lists 2003 (fw03) to assess spoken-word intelligibility in japanese," *Speech Communication*, vol. 51, no. 1, pp. 76–82, 2009.

[18] H. Scheffé, "An analysis of variance for paired comparisons," *Journal of the American Statistical Association*, vol. 47, no. 259, pp. 381–400, 1952.

[19] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[20] K. Yamamoto, T. Irino, T. Matsui, S. Araki, K. Kinoshita, and T. Nakatani, "Predicting speech intelligibility using a gammachirp envelope eistortion index based on the signal-to-distortion ratio." in *Proceedings of Interspeech 2017*, 2017, pp. 2949–2953.

[21] K. Yamamoto, T. Irino, S. Araki, K. Kinoshita, and T. Nakatani, "Gedi: Gammachirp envelope distortion index for predicting intelligibility of enhanced speech," *Speech Communication*, vol. 123, pp. 43–58, 2020.

[22] J. Hsu, *Multiple comparisons: theory and methods*. Chapman and Hall/CRC, 1996.

[23] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility." *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 436–446, 7 2013.

[24] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech Communication*, vol. 41, pp. 331–348, 2003.

[25] K. Yamamoto, T. Irino, T. Matsui, S. Araki, K. Kinoshita, and T. Nakatani, "Speech intelligibility prediction based on the envelope power spectrum model with the dynamic compressive gammachirp auditory filterbank," in *Proceedings of Interspeech 2016*, 2016, pp. 2885–2889.

[26] K. Yamamoto, T. Irino, N. Ohashi, S. Araki, K. Kinoshita, and T. Nakatani, "Multi-resolution gammachirp envelope distortion index for intelligibility prediction of noisy speech." in *Proceedings of Interspeech 2018*, 2018, pp. 1863–1867.

[27] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank." *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 2222–2232, 2006.

[28] B. C. J. Moore, *An introduction to the psychology of hearing*, 6th ed. Brill, 2013.

[29] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2009, pp. 131–137.

[30] S. Higashiyama, H. Yoshigi, H. Kawahara, and T. Irino, "Production and auditory features of speech sounds before and after vocal training using hearing impairment simulator, whis," in *Techinical Report of IEICE, Mar. 2020 (in Japanese)*.