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

Intelligibility Enhancement Based on Speech Waveform Modification Using Hearing Impairment

Shu Hikosaka[1], *Shogo Seki*[1], *Tomoki Hayashi*[1,2], *Kazuhiro Kobayashi*[1], *Kazuya Takeda*[1], *Hideki Banno*[3], *Tomoki Toda*[1]

[1]Nagoya University, Japan
[2]Human Dataware Lab. Co., Ltd., Japan
[3]Meijo University, Japan

hikosaka.shu@g.sp.m.is.nagoya-u.ac.jp, seki.shogo@g.sp.m.is.nagoya-u.ac.jp, hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp, root.4mac@gmail.com, takeda@is.nagoya-u.ac.jp, banno@meijo-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

In this paper, we propose a speech waveform modification method which incorporates a hearing impairment simulator, to improve speech intelligibility for the hearing-impaired. The settings of hearing aid devices usually need to be manually adjusted to suit the needs of each user, which creates a significant burden. To address this issue, the proposed method creates a spectral shaping filter, using a hearing impairment simulator capable of estimating speech signals as perceived by a specific hearing-impaired person. We conduct objective and subjective evaluations through simulations using the hearing impairment simulator. Our experimental results demonstrate that; 1) the proposed spectral shaping filter can significantly improve both speech intelligibility and quality, 2) the filter can be combined with a well-known speech intelligibility enhancement technique based on power compensation using dynamic range compression (DRC), and 3) speech intelligibility can be further improved by controlling the trade-off between filtering and DRC-based power compensation.

Index Terms: hearing aid, hearing impairment simulator, speech intelligibility, mel-log spectrum approximation filter, dynamic range compression

1. Introduction

Speech is one of our most important communication tools, and it allows us to communicate not only linguistic information but also paralinguistic information such as emotions and nuances. Hearing loss makes it difficult to hear the speech of others, preventing people from communicating effectively. In 2019, it was estimated that there were 14.3 million people suffering from hearing loss in Japan [1], and 466 million people worldwide [2]. Hearing loss can be divided into three categories; 1) conductive hearing loss (sound waves are not transmitted properly within the ear), 2) sensorineural hearing loss (damage to the inner ear or auditory nerve hinders the perception of sound), and 3) mixed hearing loss (a combination of conductive and sensorineural hearing loss) [3]. Sensorineural hearing loss is especially difficult to treat medically, hence hearing aid devices are typically used to assist people with this type of hearing loss.

The main function of hearing aids is to amplify the power of the input signal for each frequency band. A patient-dependent audiogram is collected, which is a graph showing a patient's audible thresholds at various standardized frequencies, to determine the degree of amplification needed for each frequency band. However, it is difficult to measure the auditory characteristics of hearing loss precisely, and therefore it is necessary to adjust the hearing aid's parameters for each patient manually during an examination. Although manual adjustment can improve intelligibility, the adjustable parameters of hearing aids are limited. Moreover, these parameters must be updated from time to time, because the auditory characteristics of the hearing-impaired are not constant, but vary over the years. This can create a significant burden for these patients, especially if progressive hearing loss goes undetected.

In this paper, we propose a novel speech modification method which incorporates a hearing impairment simulator [4, 5, 6, 7] to improve speech intelligibility, in which the differential mel-cepstrum between the original and simulated speech is filtered using a mel-log spectrum approximation (MLSA) filter [8] to compensate for the auditory characteristics of an individual's hearing loss, as modeled by the hearing impairment simulator. Since the hearing impairment simulator can precisely estimate the non-linear characteristics of hearing loss, it is expected that this filtering can compensate for these non-linear characteristics and improve the intelligibility of heard speech.

2. Hearing impairment simulator

Figure 1 shows an example of the relationship between the sound pressure levels of an input signal and its representation inside the human ear [9, 10]. The input signals are considerably amplified at low sound pressure levels, while being amplified linearly at higher pressure levels. Input signals at middle levels, i.e., 30-80 dB are more gently amplified in comparison. This pattern is known as a "compression characteristic". If the compression characteristic of a listener's hearing falls too far outside this pattern, input signals with low sound pressure levels will fail to surpass the level necessary for the perception of audible sound, and the signals will not be perceived. This is considered to be one of the major causes of hearing impairment.

A hearing impairment simulator is a system which modifies the input signal to simulate the characteristics of various types of hearing loss [4, 5, 6, 7]. By applying processing which cancels the compression characteristic and reduces sound pressure, users are able to experience how a hearing-impaired person hears sound. Since the simulator can control the level of cancellation of the compression characteristic, various types of hearing loss can be simulated. Figure 2 shows a comparison of the spectrograms of speech signals before and after applying the simulator. We can see how the signal loses its high-frequency components after applying the simulator.

Copyright © 2020 ISCA

http://dx.doi.org/10.21437/Interspeech.2020-2062

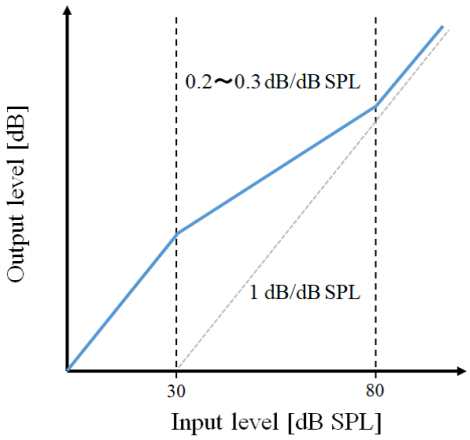

Figure 1: *An example of the input-output characteristics of the auditory periphery of people with normal hearing. Input signals between 30-80 dB are more gently amplified in comparison with other power levels.*

Many researchers have investigated the simulation of hearing loss [11]. One of the most promising approaches uses a dynamic compressive gammachirp filterbank (dcGC-FB) [12]. This approach allows us to accurately approximate not only the listener's frequency selectivity, but also the compression characteristic of their auditory periphery in comparison to listeners with normal hearing [13, 14, 15].

3. Speech enhancement using hearing impairment simulator

3.1. Enhancement using differential mel-cepstrum

Figure 3 shows a diagram of our proposed speech enhancement system, in which the auditory characteristics of hearing loss, as modeled by the hearing impairment simulator, are approximated by a differential mel-cepstrum, which is calculated as follows:

$$\mathbf{c}_d(t) = \mathbf{c}_i(t) - \mathbf{c}_o(t) \tag{1}$$

where $\mathbf{c}_i(t)$ and $\mathbf{c}_o(t)$ represent the mel-cepstrum extracted from the input and output signals of the simulator, respectively, and t represents the frame index. In this paper, we use a simulator based on the dynamic compressive gammachirp filterbank (dcGC- FB) [12]. We also investigated the effects of time-variant filtering by using the differential mel-cepstrum in each time frame, as well as time-invariant filtering using the time-averaged differential mel-cepstrum. For time-variant filtering, although it is possible to vary the enhancement filter for small speech segments, such as phonetic information, this always requires the use of the hearing impairment simulator, which complicates processing. On the other hand, for time-invariant filtering, the mel-cepstrum coefficients retain a constant value over utterances thanks to averaging processing. Therefore, it is possible to model the differential beforehand, and it is not necessary to apply the impairment simulator during testing.

In the proposed method, the input signal is filtered by the differential mel-cepstrum, based on a mel-log spectrum approximation (MLSA) filter [8] in order to enhance the components which are difficult for the hearing-impaired person to detect. Although our speech enhancement system cannot take the non-linear characteristics of hearing loss into consideration, enhanced speech is restored by adjusting the power of the input and output signals based on the differential mel-cepstrum.

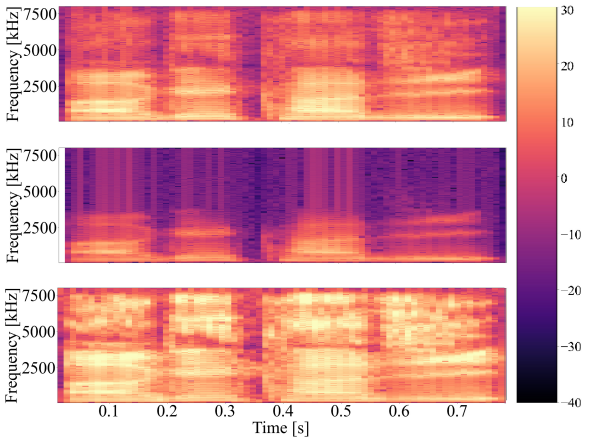

Figure 2: *Spectrograms of speech without (top) and with (center) the application of the hearing impairment simulator and a spectrogram of speech after the proposed speech enhancement system (bottom). The signal loses its high-frequency components after applying the simulator. The signal obtains its high-frequency components after applying the proposed system.*

3.2. Modification of power of filtered speech

After obtaining enhanced speech using the differential mel-cepstrum, power modification is applied as follows:

$$P_i = \frac{1}{N}\sum_{n=0}^{N} |x_i(n)|^2, \tag{2}$$

$$P_o = \frac{1}{N}\sum_{n=0}^{N} |x_o(n)|^2, \tag{3}$$

$$x_{\mathrm{mod}}(n) = \sqrt{\frac{P_i}{P_o}}\, x_f(n), \tag{4}$$

where $x_i(n)$ and $x_o(n)$ represent the input and output of the hearing impairment simulator, respectively, and n represents the time index, while $x_f(n)$ is the speech after MLSA filtering. Note that this power modification can cause an overflow. To avoid this, we set a limit to the maximum value of the coefficient in Eq. (4), so that the amplitude of the filtered signals so does not exceed a specific range. Figure 2 shows a comparison of the spectrograms of speech signals before and after applying the proposed system. We can see how the signal obtains its high-frequency components after applying the proposed system.

In this paper, we also investigate the effectiveness of dynamic range compression (DRC) [16, 17], which is used to estimate the envelope of an input signal from its amplitude. Both dynamic and static compression are then applied, based on the estimated envelope. Consequently, DRC can make the envelope of the input signal flatter, preventing overflow. DRC was inspired by compression techniques used in audio broadcasting and hearing aid amplification [18]. It is known that DRC can be used to reallocate energy in the time domain and improve speech intelligibility [16, 17].

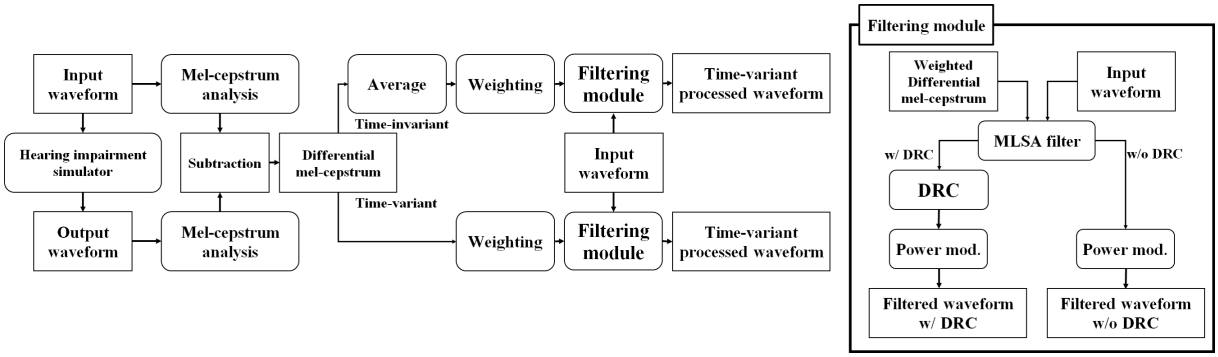

Figure 3: *A diagram of the proposed speech enhancement system.*

Table 1: *Results of objective evaluation. Mel-cepstrum distortion (MCD) and power ratio (PR) were used to compare the performance of the baseline and proposed methods.*

	MCD [dB]	PR [dB]
Baseline1	14.93	-8.55
Baseline2	15.65	-0.02
Baseline3	15.64	0.89
Proposed1	3.40	-17.36
Proposed2	3.07	-19.27

4. Experimental evaluation

4.1. Experimental conditions

The proposed method was evaluated objectively and subjectively using audio samples of words from the Familiarity-controlled Word Lists 2003 (FW03) corpus [19], a carefully designed list of Japanese words which uses "intimacy" to represent word familiarity. All of the words in the FW03 corpus are phonetically-balanced. In our experiments, we used 50 utterances of 1 male and 1 female speakers, or a total of 100 utterances. The hearing impairment simulator was used to simulate the hearing of a 60-year-old with an age-related hearing impairment, which is one type of non-linear sensorineural hearing loss. As the baseline methods, we used the output of the hearing impairment simulator without any modifications, a method with power modification without DRC, and a method with power modification with DRC. We also tested the proposed method using time-invariant and time-variant approaches. Thus, we compared the performance of the following five methods:

- **Baseline1**: No modification
- **Baseline2**: Power modification without DRC
- **Baseline3**: Power modification with DRC
- **Proposed1**: With time-invariant filter
- **Proposed2**: With time-variant filter

The sampling rate was set to 16 kHz. The sampling rate was set to 16 kHz. We used the 0th through 24th mel-cepstrum coefficients as converted from a spectral envelope, which were then analyzed using WORLD [20] (D4C edition [21]). The hearing impairment simulator was applied to the signals obtained when using each method, and the output was then evaluated.

4.2. Objective evaluation

As objective evaluation measures, we used mel-cepstrum distortion (MCD) and power ratio (PR) between the output speech of the hearing impairment simulator and the original input speech. MCD represents the difference in the spectral envelopes, and is calculated as follows:

$$\mathrm{MCD} = \frac{10}{\ln 10}\sqrt{2\sum_{d=0}^{24}(m_d^{\mathrm{proc}} - m_d^{\mathrm{ori}})^2}, \qquad (5)$$

where m_d^{proc} and m_d^{ori} represent the d-th mel-cepstrum of the output and original input speech, respectively. The smaller the MCD, the less degradation of the sound quality. On the other hand, PR represents the difference in power. The closer the PR value is to 0 dB, the more power that remains after processing.

Table 1 shows the average MCD and PR for each of the methods tested. These experimental results demonstrate that the proposed methods can significantly improve MCD compared with the baseline methods. When comparing the approaches using time-variant and time-invariant filters, the method with the time-variant filter outperformed the time-invariant filter approach. On the other hand, the PRs of both of the proposed methods were significantly lower than the baseline methods. We believe this is because the proposed method enhances the dynamic range of the waveform by filtering it with the differential mel-cepstrum, thus it is necessary to modify the range of the audio signal based on the power modification procedure described in Sec. 3.2 to avoid overflow.

Figure 4 shows the experimental results for MCD and PR when varying the weight of the differential mel-cepstrum. We varied the weight from 0 to 1 at every step of 0.1. As shown in Figure 4, there is a trade-off between the MCD and PR when using the proposed methods. We can also confirm that all of the methods tested exhibited the same tendencies. On the other hand, applying DRC to enhance the power of the enhanced speech tended to reduce power degradation while maintaining a lower level of MCD. Based on these objective results, it can be said that the proposed methods with DRC achieved better performance compared to the other methods when enhancing speech to compensate for hearing loss.

4.3. Subjective evaluation

We also conducted a subjective evaluation using written listening tests. Subjects using headphones listened to word utterances of 4 mora as output by the hearing impairment simulator, and were asked to transcribe the content. After transcribing each

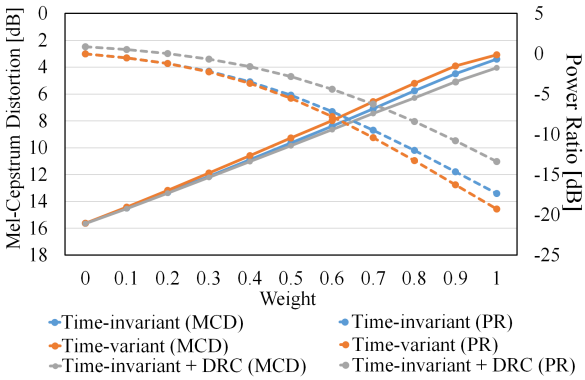

Figure 4: *Relationship between mel-cepstrum weight and objective evaluation scores.*

Table 2: *Experimental results for written test.*

Method	Accuracy [%]	MOS
Baseline1	39.38	3.13
Baseline2	50.00	3.43
Baseline3	52.50	3.29
Proposed1	54.38	3.63
Proposed2	55.63	3.74

utterance, subjects were asked to provide a confidence level for their answer using a five-level opinion rating (1: low level of confidence, 2: somewhat low level, 3: not a high level, 4: somewhat high level, and 5: high level). Each subject listened to 50 word utterances by speakers of each gender, for a total of 100 utterances. Our subjects consisted of eight males in their 20s. By having our subjects listen to the output of the hearing impairment simulator, our goal was to simulate the subjective evaluations of hearing-impaired persons.

Table 2 shows the average accuracy rates for the written test and average Mean Opinion Scores (MOS) related to the subjects' confidence levels in their answers. The higher the MOS, the more confident the subjects were in their answers. Applying power modification in the Baseline2 method drastically improved the accuracy of word recognition when compared to the Baseline1 method without processing. Our proposed methods achieved further improvements in word recognition compared to the baseline methods. We can also see that the proposed method with time-invariant filtering achieved comparable performance to the proposed method with time-invariant filtering. This result implies that the sophisticated modeling which is applied when using the time-variant filter is not more effective for improving speech intelligibility.

We conducted further subjective evaluations to assess the performance of the proposed method when using DRC. Table 3 shows the experimental results for accuracy and MOS at each weight setting. These results demonstrate that using DRC with the proposed method results in about a 6 % improvement in word accuracy compared to the proposed method without DRC. Moreover, by adjusting the mel-cepstrum weight, we were able to achieve an additional improvement in word accuracy of more than 6 %.

Table 3: *Experimental results for written test for proposed method using time-invariant filtering with DRC at various mel-cepstrum weights.*

Weight	Accuracy [%]	MOS
0.0	52.50	3.29
0.6	55.71	3.47
0.8	68.57	3.64
1.0	62.14	3.58

5. Conclusion

In this paper, we proposed a novel speech enhancement technique which incorporates a hearing impairment simulator for improving the intelligibility of heard speech. Our experimental results demonstrated that the proposed method makes it possible to significantly improve the sound quality of enhanced speech while reducing speech power in regard to the "compression characteristic". Furthermore, we found that DRC is an effective method for amplifying reduced power, and that speech intelligibility can be further improved by adjusting a filtering weight. The proposed method with a time-invariant filter and DRC achieved an approximately 18 % relative improvement in word recognition accuracy compared to the conventional, simple power modification method used in our subjective evaluation.

For future work, we plan to experiment with filter design by exploring methods of clipping processing, while utilizing an objective index of speech intelligibility for evaluation. Furthermore, we plan to experiment with filter design considering the effect of noise.

6. Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Number 16H01734 and JST, CREST Grant Number JPMJCR19A3.

7. References

[1] Anovum, "Japan Trak 2018," Japan Hearing Instruments Manufacturers Association, Tech. Rep., 2018.

[2] W. H. Organization and I. T. Union, "Safe listening devices and systems: a WHO-ITU standard," Tech. Rep., 2019.

[3] R. Smith, A. Shearer, M. Hildebrand, and G. Camp, *Deafness and Hereditary Hearing Loss Overview*, 2008.

[4] E. Villchur, "The effect of recruitment on speech perception—a simulation for normal listeners," *The Journal of the Acoustical Society of America*, vol. 55, no. 2, pp. 450–450, 1974.

[5] 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: Physiology and Perception*, 2013.

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

[7] M. Nagae, T. Irino, R. Nisimura, and H. Kawahara, "Processing of inverse compression and user-interface for hearing impairment simulator," *Proceedings of the auditory research meeting*, vol. 44, pp. 13–18, 2014.

[8] S. Imai, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *IEICE Transactions on Fundamentals*, vol. 66, no. 2, pp. 122–129, 1983.

[9] T. Irino, "Invited lecture: Measurement and formulation of the auditory filter," *Proceedings of the Auditory Research Meeting*, vol. 39, no. 6, pp. 413–418, 2009.

[10] T. Irino, "Psychophysical measurement of cochlear compression and its application to a hearing impairment simulator," *Proceedings of the Acoustical Society of Japan*, pp. 1579–1582, 2014.

[11] Y. Nejime and B. C. J. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 603–615, 1997.

[12] T. Matsui, H. Banno, R. Nishimura, and T. Irino, "Implementation of hearing impairment simulator based on the gammachirp auditory filterbank and its educational application," *IEICE Technical Report*, vol. 118, no. 269, pp. 31–36, 2018.

[13] T. Irino and R. D. Patterson, "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] R. D. Patterson, M. Unoki, and T. Irino, "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *The Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1529–1542, 2003.

[15] 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.

[16] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *13th Annual Conference of the International Speech Communication Association*, 2012, pp. 634–637.

[17] T.-C. Zorila and Y. Stylianou, "On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2050–2054, 01 2014.

[18] B. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 1, pp. 22–32, March 1969.

[19] 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.

[20] 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. 99, no. 7, pp. 1877–1884, 2016.

[21] M. Morise, "D4c, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.