Journal of the Institute of Industrial Applications Engineers Vol.9, No.3, pp.97–101, (2021.7.25)
DOI: 10.12792/JIIAE.9.97 Online edition: ISSN 2187-8811 Print edition: ISSN 2188-1758

Paper

Power-saving Reproduction Algorithm for Speech Signals

MITSUHIRO NAKAGAWARA*†‡ Non-member, MITSUNORI MIZUMACHI‡ Member

(Received May 27, 2021, revised July 15, 2021)

Abstract: Power consumption is one of the important issues for portable electronic devices. The authors have previously proposed a power-saving audio reproduction algorithm based on auditory masking. In this paper, the feasibility of the power-saving audio playback algorithm is confirmed towards speech signals in the viewpoints of both speech intelligibility and power consumption. The influence of the speech conversion for power-saving playback was quantified as the word intelligibility. In practice, the word intelligibility was measured in each of six phoneme categories for the converted speech signals of which the reduction rate of power consumption was set at 20%, 50%, and 70%. It is confirmed that the word intelligibility does not decrease drastically for the converted speech of which the reduction rate of power consumption was set at 50%.

Keywords: Power-saving, Speech signal, Word intelligibility,

1. Introduction

Due to the influence of the COVID-19 that has raged around the world, an online meeting using smartphones and tablet devices has been increased. However, portable devices require batteries. Power-saving audio reproduction has been studied both in hardware and software implementation. The hardware issues include the design of an amplifier [1] and a digital-to-analog converter [2]. Recently, without using any analog components, a fully digital audio system that uses digital loudspeakers with multiple coils achieves energy saving [3] [4]. These attempts can efficiently reduce power consumption, but it is necessary to redesign the audio system. On the other hand, the software attempts are relatively easier to be implemented with digital signal processing. For example, it is achieved by analyzing an input sound source in real-time and reducing the amplitude signal levels in the temporal domain [5] [6]. However, the amplitude suppression in the temporal domain might cause the deterioration of sound quality.

The authors have approached a software-oriented power-saving audio reproduction based on auditory characteristics which are implemented with the filterbank process [7] and the masking processing [8]. First, the proposed method focused on the musical source and achieved the reduction of power consumption by 25% [7]. Then, the effectiveness of the proposed method was confirmed using a variety of music sources in simulated noisy conditions. The results of the listening tests suggested that the sound quality was mostly satisfactory under noisy environments [9].

In this study, power-saving audio reproduction is investigated by focusing on speech signals. It is important how accurately we can understand meaningful words when we perceive speech [10]. Compared to music sources, much power-saving can be achieved for speech signals unless the word intelligibility does not deteriorate. In this paper, the relationship between the word intelligibility and current consumption is investigated by carrying out the diagnostic rhyme test in Japanese.

In Sec. 2, the proposed power-saving reproduction algorithm is explained. Sec. 3 describes the measuring method of current consumption while reproducing speech signals. Sec. 4 describes the procedure of a listening test for the subjective evaluation, and results, followed by the conclusion in Sec. 5.

2. Power-saving Reproduction Algorithm

2.1 Outline The overview of the proposed power-saving audio reproduction algorithm is summarized in Fig. 1. The method mainly consists of the auditory-oriented sub-band optimization in the filterbank processing and the reduction of inaudible components below the masking thresholds.

2.2 Auditory Masking Auditory masking occurs when the perception of one sound is affected by the presence of another sound. For example, as shown in Fig. 2, the frequency component with a low sound pressure level (A) located near a high sound pressure level (B) is masked and cannot be perceived. The concept of auditory masking is widely used in auditory-oriented signal processing. The MPEG audio codec has been a great success in lossy audio coding [8]. It is supposed that the power-saving audio reproduction can be further improved by employing the auditory masking based on the MPEG-1 model [11] which can suppress the perceptual distortion. It aims to calculate the masking threshold level of the perceivable limit from the curve between the original sound source and the minimum

* Corresponding: nakagawara.mitsuhiro@jp.panasonic.com
† Panasonic Corporation
4261 Ikonobe-cho, Tsuzuki-ku, Yokohama-shi, Kanagawa, Japan 224-8520
‡ Kyushu Institute of Technology
1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka, Japan 804-8550

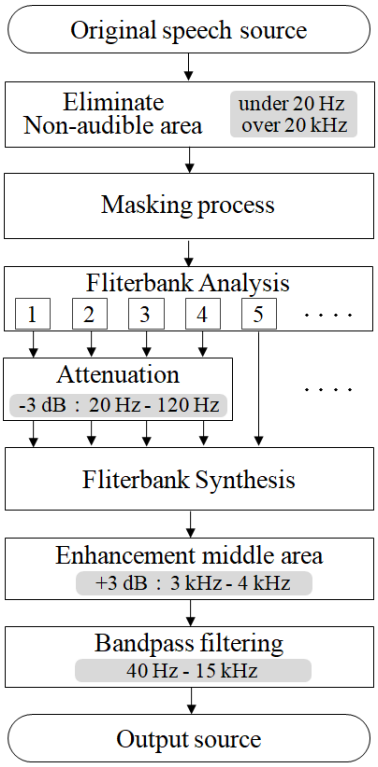

Figure 1: Proposed power-saving playback algorithm.

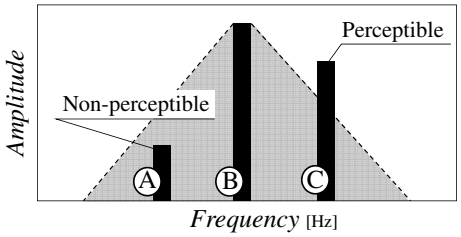

Figure 2: Concept of the auditory masking.

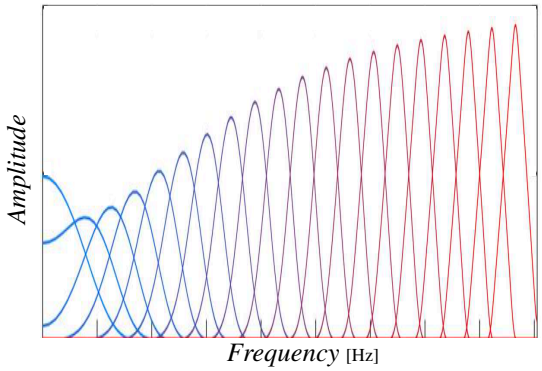

Figure 3: Mel-frequency filterbank.

audible levels.

In this study, the reduction levels below the masking threshold were prepared on four attenuation patterns of 3 dB, 10 dB, 20 dB, and 30 dB compared with the original sound, and 10 dB was adopted, which gave the least deterioration in the sound quality.

2.3 Auditory Filterbank The human performs frequency analysis of sound by mechanical vibration of the basilar membrane in the cochlea. The auditory filterbank, which arranged the different center frequency bands, approximates the behavior of the basilar membrane [12] [13].

In this study, the mel-frequency filterbank, which arranges triangular windows at equal intervals on the mel-frequency scale, is employed because it is used for the front-end of automatic speech recognition systems [14]. It approximates the coordinates on the basilar membrane where the lower frequency is more emphasized as shown in Fig. 3.

2.4 Implementation The proposed method relies on auditory characteristics for eliminating non-audible frequency components below 20 Hz and upper 20 kHz. The auditory masking occurs when the perception of the target sound is affected by the presence of another sound. The masking threshold is calculated by using the psychoacoustic model used in the MPEG-1 codec [11]. Then, the frequency components below the masking threshold are eliminated before the filterbank analysis [8].

After the filterbank analysis and synthesis, the sub-band power is decreased at 3 dB at 120 Hz and lower frequency components. It was confirmed that the current consumption was suppressed by about 30% when the lower frequency components of an audio source were dropped by using a low-pass filter, Therefore, it is essential to reduce the lower frequency energy.

Finally, band-pass filtering in between 40 Hz to 15 kHz was applied for cutting out the extra high and low-frequency components.

When the proposed method deals with speech signals, the reduction ratios of current consumption are adjusted at approximately 20%, 50%, and 70% by setting the number of filterbanks at 5, 15, and 25 where the whole frequency range of the filterbank is fixed up to 20 kHz, respectively. The word intelligibility obtained by performing a listening test is measured for each adjusted current consumption.

3. Measurement of Power Consumption

Power consumption was measured using a circuit shown in Fig. 4. It simulated a typical audio output configuration consisting of a DSP to a D/A converter and a power amplifier. The prepared speech signals were played back through the D/A converter (Marantz HD-DAC1), the power amplifier (YAMAHA P4050), and a loudspeaker (Fostex FE83En). The current consumption was obtained from the voltage in between the resister, and it was averaged out over 5 measurements to minimize the measurement error. The amplitude levels of the speech signal were normalized by the maximum value.

The speech signals were prepared from the male utterances recorded in a soundproof room. The utterances were prepared from the word pair lists, which will be described in Sec. 4. The audio formats were in WAV (16 bits, 44.1 kHz).

Figure 5 shows the averaged reduction ratios for each version of the filterbanks with 5, 15, and 25 channels, re-

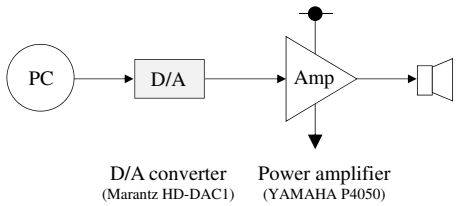

Figure 4: Measurement circuits of power consumption.

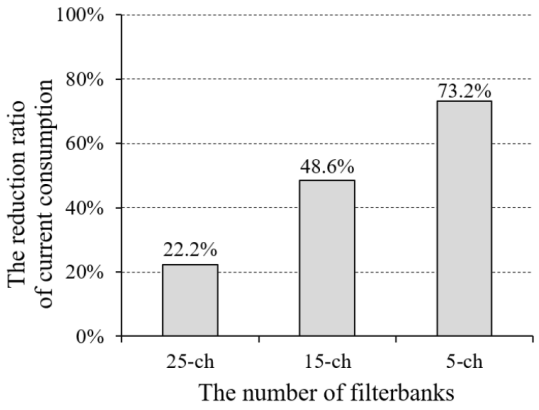

Figure 5: The reduction ratio of current consumption with the number of filterbanks for proposed algorithm.

Table 1: The phenome categories.

No.	Phenomes	Corresponding classification
1	Voicing	vocalic-nonvocalic
2	Nasality	nasal-oral
3	Sustention	continuant-interrupted
4	Sibilation	strident-mellow
5	Graveness	grave-acute
6	Compactness	compact-diffuse

spectively. As mentioned in Sec. 2.4, the current consumption reduction ratios depend on the number of the filterbank channels. A higher reduction of the current consumption is achieved with less number of the channels.

4. Diagnostic Rhyme Test

It is important for our speech communication to accurately convey the contents of the utterances rather than the sound quality. In general, speech intelligibility can be measured with phonemes, words, and sentences [15]. In this study, the word intelligibility is measured by carrying out the Diagnostic Rhyme Test (DRT) standardized by ANSI [16]. In the DRT, two words differ only in their initial, and the two consonants differ only in a single distinctive acoustic phonetic feature. The intelligibility of each phoneme feature can be evaluated. Therefore, the word intelligibly of each phoneme feature can be evaluated. The six phoneme categories, which were defined by Jacobson *et al.*, were used as shown in Table 1. The Japanese DRT word pair lists shown in Appendix A [17] were used in the experiment. It is confirmed that the six phoneme categories are suitable for evaluating speech intelligibility in real acoustic environments [18]. In each word pair, the first phonemes are different and

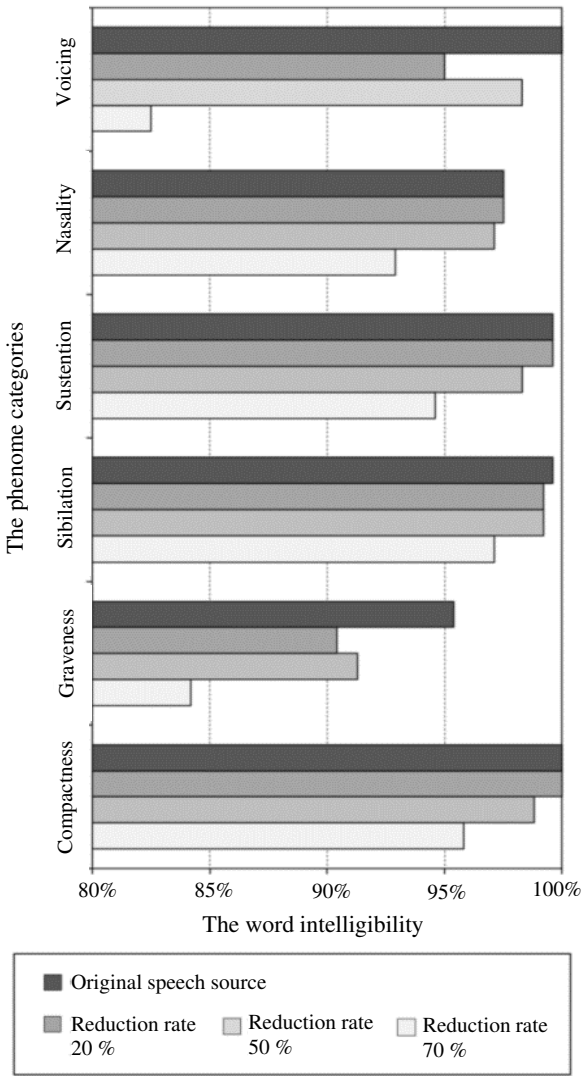

Figure 6: The word intelligibility of each phoneme with the reduction ratio of current consumption.

the latter phonemes are the same. The subjects listened to the above words and were asked to write down what they have heard. The validity is judged by whether the meaning of the word can be heard correctly.

Subjects were 24 university students, who were 6 females and 18 males, with normal hearing. The DRT was carried out in the soundproof room of Kyushu Institute of Technology. The listening tests was performed as follows.

(1) A word was randomly selected from the ten pairs of Japanese DRT words to be presented to the subject.

(2) After listening the presented word, the subjects wrote down what they heard in the answer sheet shown in Appendix B in three seconds.

(3) The next word was automatically presented to the subjects after the three seconds break. It continued until the whole words of the ten pairs were presented.

(4) The above steps: (1), (2), and (3) were repeated in each phoneme category.

Figure 6 shows the results of the DRT in each phoneme category for the original speech signals and the processed speech signals of which the power consumption reduction rates are 20%, 50%, and 70%, respectively. It is confirmed that the word intelligibility rates over 80% were obtained in all phoneme categories even in the case of the reduction rate of 70%. It is necessary to perform DRT with a wide variety of participants whose ages range widely.

5. Conclusions

The power-saving audio playback algorithm is modified for speech reproduction systems. The number of mel-frequency filterbanks was adjusted to achieve the reduction of power consumption up to approximately 20%, 50%, and 70%. The feasibility of the modified algorithm was confirmed by carrying out the Japanese Diagnostic Rhyme Test in six phoneme categories. The experimental results suggest that the word intelligibility rates attain 90% and more in the whole phoneme categories for the processed speech signals with the reduction rates of 20% and 50%. Even for those with the reduction rate of 70%, the intelligibility rate exceeds 80%. It is indicated that the power consumption can be significantly decreased in the proposed method by smartphones and tablet devices. It is also useful for the situation of an online meeting which has been glowing. Future works include the performance evaluation under practical conditions. Future works include the additional listening test for investigating the quality of the processed speech in detail.

References

[1] P. Tournier, "Class AB versus D for Multimedia Applications in Portable Electronic Devices", *AES 29th International Conference: Audio for Mobile and Handheld Devices*, Paper No.2-2, 2006.

[2] B. Trotter, J. Tucker and J. Rhode, "Low-Voltage, Low-Power Converter Design for Portable Audio Applications", *AES UK 16th Conference: Silicon for Audio*, Paper No.uk037, 2001.

[3] J. Okamura and A. Yasuda, "Digital speaker driving apparatus", U.S. Patent No.8, 306, 244, 6 Nov. 2012.

[4] Y. Furuya, M. Takahashi, S. Saikatsu, M. Yoshino and A. Yasuda, "Speaker system with 100-W high output power and 0.17% THD using 9-V power supply with digitally direct-driven technique", *2017 IEEE International Conference on Consumer Electronics (ICCE)*, pp.452-455, 2017. DOI: 10.1109/ICCE.2017.7889393

[5] T. Ishizuka and Y. Aso, "Sound amplifying apparatus", *Japanese Unexamined Patent Application Publication*, No.2009-253955, 29 Oct. 2009. (in Japanese)

[6] M. Yoneda, K. Ookuri and A. Kakema, "Audio signal processing device for adjusting volume", *U.S. Patent*, No.9, 667, 213, 30 May 2017.

[7] M. Mizumachi, W. Kubota and M. Nakagawara, "Power Saving Audio Playback Algorithm Based on Auditory Characteristics", *Proc. 144th AES Conv.*, Paper No. 9933, 2018.

[8] T. Nakashima, M. Nakagawara and M. Mizumachi, "Proposal of Power-saving Audio Playback Algorithm Based on Auditory Masking", *Proc. 146th AES Conv.*, Paper No.10164, 2019.

[9] T. Nakashima, M. Nakagawara and M. Mizumachi, "Subjective evaluation of power saving audio playback algorithm based on auditory masking", *IEICE Technical Report*, 2019.

[10] Acoustical Society of Japan, "Introduction to Acoustics", *Corona publishing*, 2011. (in Japanese) DOI: 10.1007/978-1-4939-0755-7_1

[11] ISO/IEC 11172-3, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s-Part 3: Audio", 1993.

[12] T. Irino, "An introduction to auditory filter", *Journal of the Acoustical Society of Japan*, 66(10), pp.506-512, 2010. (in Japanese)

[13] Acoustical Society of Japan (2011), "Models in Hearing", *Corona publishing*, pp.102-128, 2011. (in Japanese)

[14] D.O'Shaughnessy, "Speech communication: human and machine", *Addison-Wesley*, 1987.

[15] Acoustical Society of Japan, "Acoustic Keyword Book", *Corona publishing*, 2016. (in Japanese)

[16] ANSI Standard S3.2-1989, "Method for Measuring the Intelligibility of Speech over Communication Systems", (1989, reaffirmed 1995).

[17] K. Kondo, R. Izumi, M. Fujimori, R. Kaga and K. Nakagawa, "On a Two-to-one Selection Based Japanese Speech Intelligibility Test", *Journal of the Acoustical Society of Japan*, 63(4), pp.196-204, 2007. (in Japanese)

[18] M. Fujimori, K. Kondo, K. Takano and K. Nakagawa, "On a revised word-pair list for the Japanese intelligibility test", *Proc. International Symposium on Frontiers in Speech and Hearing Research*, H-2006-19, March 2006.

Nakagawara Mitsuhiro (Non-member) received the B.S. and M.S. degrees in Electrical Engineering from Kyushu Institute of Technology in 2008 and 2010, respectively. Since 2010, he has been working for Panasonic Corporation. He is engaged in the design of car navigation systems.

Mitsunori Mizumachi (Member) received the Bachelor degree in design from Kyushu Institute of Design and the Ph.D degree in information science from Japan Advanced Institute of Science and Technology (JAIST) in 1995 and 2000, respectively. He is currently an associate professor at Kyushu Institute of technology. His research interests include acoustic information processing and statistical signal processing. He is a member of AES, ASA, ASJ, IEEE, IEICE, and RISP.

Appendix

A. Japanese DRT words pair lists

Figure A.1 shows the Japanese DRT words pair lists.

B. Answer sheet of listening tests

Figure B.1 shows the answer sheet of listening tests

Voicing	Nasality	Sustention	Sibilation	Graveness	Compactness
財 - 才 （ザイ - サイ）	万 - 番 （マン - バン）	箸 - 菓子 （ハシ - カシ）	ジャム - ガム （ジャム - ガム）	枠 - 楽 （ワク - ラク）	焼く - 沸く （ヤク - ワク）
抱く - 炊く （ダク - タク）	無い - 台 （ナイ - ダイ）	旗 - 型 （ハタ - カタ）	着 - 角 （チャク - カク）	パイ - タイ （パイ - タイ）	貝 - パイ （カイ - パイ）
議事 - 記事 （ギジ - キジ）	ミス - ビス （ミス - ビス）	私利 - 地理 （シリ - チリ）	式 - 引き （シキ - ヒキ）	見栄 - 煮え （ミエ - ニエ）	銀 - 瓶 （ギン - ビン）
銀 - 金 （ギジ - キジ）	見る - ビル （ミル - ビル）	昼 - 着る （ヒル - キル）	知事 - 記事 （チジ - キジ）	ミス - ニス （ミス - ニス）	気障 - ピザ （キザ - ピザ）
髄 - 粋 （ズイ - スイ）	無理 - 鰤 （ムリ - ブリ）	好き - 月 （スキ - ツキ）	中 - 空 （チュー - クー）	剥く - 抜く （ムク - ヌク）	黒 - プロ （クロ - プロ）
ぐう - 食う （グー - クー）	無視 - 武士 （ムシ - ブシ）	砂 - 綱 （スナ - ツナ）	純 - 群 （ジュン - グン）	無視 - 主 （ムシ - ヌシ）	タ - ルー （ユー - ルー）
税 - 生 （ゼイ - セイ）	面 - 弁 （メン - ベン）	変 - 剣 （ヘン - ケン）	シェア - ヘア （シェア - ヘア）	面 - 年 （メン - ネン）	弦 - 弁 （ゲン - ベン）
出歯 - 手羽 （デバ - テバ）	練る - 出る （ネル - デル）	縁 - けり （ヘリ - ケリ）	シェル - 経る （シェル - ヘル）	ペン - 天 （ペン - テン）	剣 - ペン （ケン - ペン）
像 - 僧 （ゾー - ソー）	門 - 盆 （モン - ボン）	星 - 腰 （ホシ - コシ）	条 - 号 （ジョー - ゴー）	毛 - 脳 （モー - ノー）	語気 - 簿記 （ゴキ - ボキ）
誤字 - 孤児 （ゴジ - コジ）	野良 - 銅鑼 （ノラ - ドラ）	掘る - 凝る （ホル - コル）	所持 - 保持 （ショジ - ホジ）	ポロ - とろ （ポロ - トロ）	余暇 - ろ過 （ヨカ - ロカ）

Figure A.1: The Japanese DRT words pair lists.

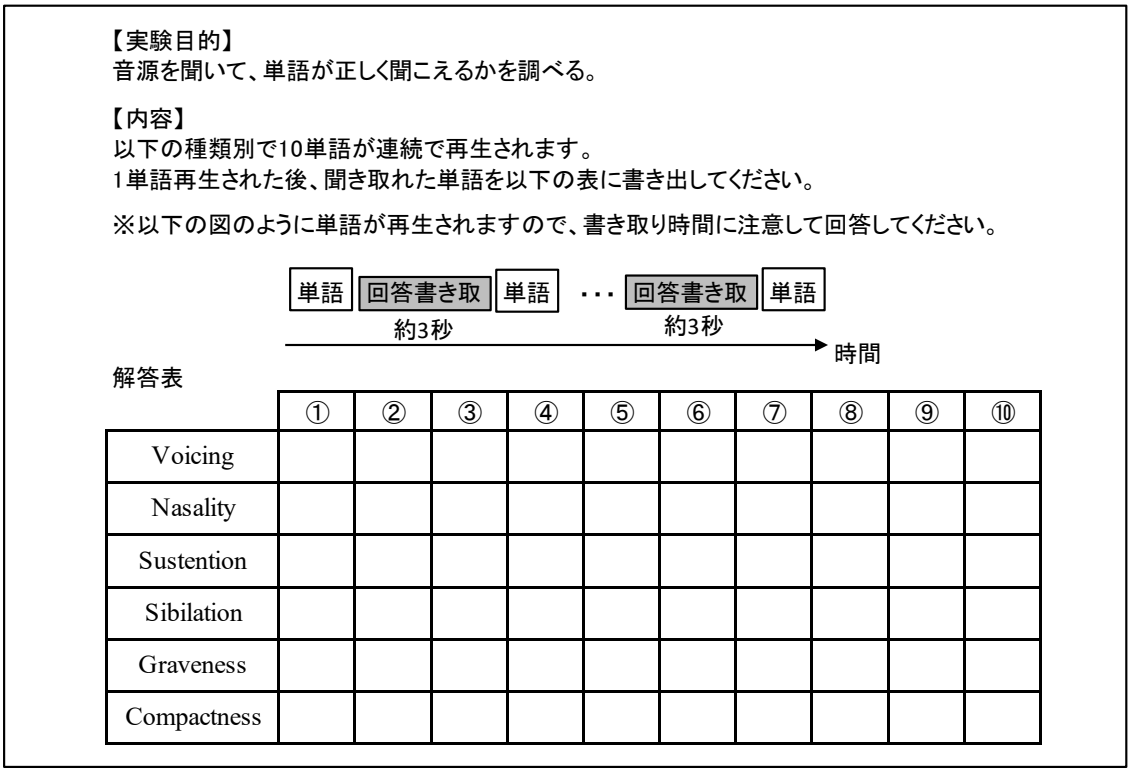

Figure B.1: The answer sheet of listening test.