INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Impact of Residual Noise and Artifacts in Speech Enhancement Errors on Intelligibility of Human and Machine

Shoko Araki[1], *Ayako Yamamoto*[2], *Tsubasa Ochiai*[1], *Kenichi Arai*[1] ,
Atsunori Ogawa[1], *Tomohiro Nakatani*[1] *and Toshio Irino*[2]

[1]NTT Corporation, Japan
[2]Faculty of Systems Engineering, Wakayama University, Japan

shoko.araki.pu@hco.ntt.co.jp

Abstract

Single-channel (1-ch) speech enhancement (SE) has been widely studied, and high accuracy has been achieved recently. However, enhanced speech still includes some errors that affect human hearing quality and SE applications, e.g., automatic speech recognition (ASR). Previously, [Iwamoto et al., "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR" in Interspeech 2022, pp. 5418–5422] decomposed the errors in an enhanced signal into residual noise and artifact components, and analyzed their impacts on ASR performance. They showed that the artifacts have a greater impact than the residual noise on ASR. Although the impact on human intelligibility has not been investigated yet, it is essential to get the knowledge to develop SE techniques suitable for both humans and machines. This paper, therefore, investigates the effects of such error factors on human listening. Our subjective test results show that the artifacts have a large impact on human intelligibility, and that residual noise has a lesser impact. This tendency is similar in machine ASR.

Index Terms: Single-channel speech enhancement, artifact, residual noise, speech intelligibility, speech recognition

1. Introduction

Single-channel (1-ch) speech enhancement (SE) is important in many applications where multiple microphones are unavailable. Recently 1-ch SE using deep learning (DL) has been widely studied and has achieved high performance [1, 2, 3, 4, 5, 6, 7]. However, the enhanced speech by 1-ch SE still contains errors due to residual noise and artifacts in enhanced speech. These errors affect both human listening and speech applications of SE, e.g., automatic speech recognition (ASR). In this paper, we investigate the impacts of residual noise and artifact components in an enhanced speech on human hearing quality and ASR. Our motivation and ultimate goal are to obtain basic data for: (a) developing SE techniques suitable for both humans and machines (including ASR) and (b) developing a rational ASR-based intelligibility prediction method.

Although subjective evaluations of 1-ch SE have recently been conducted in several challenges, e.g., [7], the impact of SE errors on subjective evaluation scores remains inadequately investigated. Venkataramani et al. [8] conducted a listening test on enhanced signals in a speech separation task using a DL-based 1-ch SE with performance measures SDR/SIR/SAR[1] in BSS_Eval [9] and STOI [10]. However, they have not reported the effects of residual noise (interference) and artifacts alone on human listening. Yamamoto et al. [11] evaluated the intelligibility of enhanced speech signals with an ideal ratio mask (IRM), which is an oracle mask widely used as a training target of DL-based SE in noise reduction tasks (e.g., [1, 2, 4]). However, they did not investigate the impact of residual noise and artifacts in enhanced speech signals on human intelligibility.

There have been many works on ASR with a 1-ch SE frontend [12, 13, 14, 15], and it had been reported that most 1-ch SE approaches tend only slightly to improve or even degrade ASR performance. Another work [16] decomposed SE errors into residual noise and artifacts of the enhanced speech, and analyzed the factors that impact ASR performance, concluding that artifacts affect ASR significantly more than residual noise. While it is important to create such SE techniques that are suitable for ASR, it is also important to consider whether these techniques are suitable for human listening as well. Furthermore, in recent years, researchers have studied using ASR as a metric to predict speech intelligibility for humans [17, 18, 19, 20]. To develop a rational predictor based on ASR, we must confirm whether human perception and ASR exhibit similar behavior to residual noise and artifacts.

This paper conducted subjective tests and investigated the impact of residual noise and artifacts in enhanced speech signals on human intelligibility. As SE, we employed the ideal ratio mask (IRM), which is widely used as a training target of DL-based SE approaches (e.g., [1, 2, 4]). We also compared the impact on human intelligibility with that on ASR systems. We report that the artifacts in SE error largely impact human intelligibility, whereas the residual noise component has a lesser impact, which is a result that similar to the impact of SE on ASR.

2. Errors in enhanced signals

In this paper, we evaluate the impact of SE errors on both human listening and ASR performance in the following procedures:

1. Obtaining enhanced signals using IRM [Sec. 2.1].
2. Decomposing SE errors into several factors [Sec. 2.2].
3. Modifying the enhanced signals by rescaling each error factor [Sec. 2.3]
4. Conducting subjective and objective tests, i.e., evaluating the modified enhanced signals in terms of intelligibility (subjective) and WER (objective) [Secs. 3 and 4]
5. Analyzing the impacts of the error factors on the intelligibility and the WER [Sec. 4].

This section explains steps 1–3.

2.1. Speech enhancement with ideal ratio mask (IRM)

This paper focuses on the 1-ch SE (noise reduction) task. Let $\mathbf{x} \in \mathbb{R}^T$ denote a T-length time-domain waveform of the ob-

[1] Signal-to-distortion ratio, signal-to-interference ratio, and signal-to-artifact ratio, respectively.

10.21437/Interspeech.2023-1116

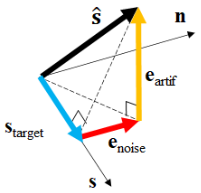

$\widehat{\mathbf{s}}$
$\mathbf{n}$
$\mathbf{e}_{\text{artif}}$
$\mathbf{s}_{\text{target}}$
$\mathbf{e}_{\text{noise}}$
$\mathbf{s}$

Figure 1: *Decomposition of speech enhancement errors.*

served signal. The observed signal is modeled as

$$\mathbf{x} = \mathbf{h} * \mathbf{c} + \mathbf{n} = \mathbf{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{c} \in \mathbb{R}^T$, $\mathbf{h} \in \mathbb{R}^L$, and $\mathbf{n} \in \mathbb{R}^T$ are a clean speech signal, an impulse response of L-tap from the speech source to a microphone, and a noise signal, respectively. "*" denotes the linear convolution, and $\mathbf{s} = \mathbf{h} * \mathbf{c}$ is a source image. The aim of SE is to reduce noise signal $\mathbf{n}$ from observed signal $\mathbf{x}$.

As a 1-ch SE, this paper employs an IRM. An IRM itself, or enhanced speech with an IRM, is widely used as a training target of deep neural networks (DNNs) for SE (e.g., [1, 2, 4]). We evaluated the original IRM [1] to ascertain the baseline of the performance. First, we obtained the time-frequency representation of Eq. (1) with a short-time Fourier transform (STFT):

$$x_{\tau f} = s_{\tau f} + n_{\tau f}, \tag{2}$$

where τ and f are time and frequency indices, respectively. The IRM for enhancing the speech signals is given as [1]

$$M_{\tau f} = \left(\frac{|s_{\tau f}|^2}{|s_{\tau f}|^2 + |n_{\tau f}|^2} \right)^{0.5}, \tag{3}$$

and the enhanced signal in the STFT domain with the IRM is

$$\widehat{s}_{\tau f} = M_{\tau f} x_{\tau f}. \tag{4}$$

Enhanced speech signal $\widehat{\mathbf{s}}$ is obtained by applying the inverse STFT and the overlap-add to $\widehat{s}_{\tau f}$.

2.2. Speech enhancement error decomposition

We employed the orthogonal projection-based error decomposition approach that was originally introduced for designing the SE evaluation metrics [9]. The estimated signal $\widehat{\mathbf{s}}$, which inevitably contains estimation errors, is decomposed into three terms: [2]

$$\widehat{\mathbf{s}} = \mathbf{s}_{\text{target}} + \mathbf{e}_{\text{noise}} + \mathbf{e}_{\text{artif}}, \tag{5}$$

where $\mathbf{s}_{\text{target}} \in \mathbb{R}^T$ is the target source component, which originated from $\mathbf{s}$, and $\mathbf{e}_{\text{noise}} \in \mathbb{R}^T$ and $\mathbf{e}_{\text{artif}} \in \mathbb{R}^T$ denote noise and artifact error components, respectively. Figure 1 illustrates this signal decomposition framework.

The decomposed terms in Eq. (5) are obtained using projection matrices:

$$\mathbf{s}_{\text{target}} = \mathbf{P}_{\mathbf{s}} \widehat{\mathbf{s}}, \tag{6}$$

$$\mathbf{e}_{\text{noise}} = \mathbf{P}_{\mathbf{s},\mathbf{n}} \widehat{\mathbf{s}} - \mathbf{P}_{\mathbf{s}} \widehat{\mathbf{s}}, \tag{7}$$

$$\mathbf{e}_{\text{artif}} = \widehat{\mathbf{s}} - \mathbf{P}_{\mathbf{s},\mathbf{n}} \widehat{\mathbf{s}}, \tag{8}$$

where $\mathbf{P}_{\mathbf{s}} \in \mathbb{R}^{T \times T}$ is the orthogonal projection matrix onto the subspace spanned by the source signals, and $\mathbf{P}_{\mathbf{s},\mathbf{n}} \in \mathbb{R}^{T \times T}$ is the orthogonal projection matrix onto the subspace spanned by the source and noise signals. See [9] and [16] for more details about projection matrices.

2.3. Modifying enhanced signals by controlling errors

After decomposing enhanced signal $\widehat{\mathbf{s}}$ with orthogonal projection-based decomposition, as in Sec. 2.2, we synthesized a modified version of enhanced signal $\widehat{\mathbf{s}}_\omega \in \mathbb{R}^T$ by directly scaling error components $\mathbf{e}_{\text{noise}}, \mathbf{e}_{\text{artif}}$:

$$\widehat{\mathbf{s}}_\omega = \mathbf{s}_{\text{target}} + \omega_{\text{n}}\, \mathbf{e}_{\text{noise}} + \omega_{\text{a}}\, \mathbf{e}_{\text{artif}}, \tag{9}$$

where ω_{n} and ω_{a} are the parameters that control the amount of noise and artifact error components.

A previous work [16] also showed that adding a scaled version of an observed signal to an enhanced signal also improved the ASR performance. We call this technique "observation adding". The procedure for the observation adding is

$$\widehat{\mathbf{s}}_\omega = (1 - \omega_{\text{oa}})\widehat{\mathbf{s}} + \omega_{\text{oa}} \mathbf{x}, \tag{10}$$

where ω_{oa} controls the amount of added observed signal to the enhanced signal. In this paper, we set $\omega_{\text{oa}} = 0.5$.

3. Experimental conditions

3.1. Noisy observations

For the clean speech $\mathbf{c}$ in Eq. (1), we used Japanese 4-mora words provided in a database, FW07, which was developed for intelligibility tests with careful control of word familiarity and phonetic balance [21]. The dataset contains 400 words for each of the four familiarity ranks, and the average duration of a 4-mora word is approximately 700 ms. The speech sources were obtained from a word set of the least familiarity to prevent increment to speech intelligibility scores by guessing commonly used words. In this paper, we used the speech sources uttered by a male speaker (label ID: mis).

We also prepared recorded noise signals $\mathbf{n}$ and impulse responses $\mathbf{h}$ in Eq. (1). For noise $\mathbf{n}$, we used recorded babble noise: multi-speaker speech signals were played simultaneously from many loudspeakers on a large office floor and recorded using a microphone in a conference room adjacent to the office floor (reverberation time: approx. 360 ms; door open). We also measured impulse response $\mathbf{h}$ from 12 loudspeaker positions to the microphone in the conference room.

Observed signal $\mathbf{x}$ in Eq. (1) was produced by convolving a clean speech $\mathbf{c}$ and an impulse response $\mathbf{h}$, which was randomly selected from 12 source positions, and adding babble noise $\mathbf{n}$. The input signal-to-noise ratio (iSNR) conditions ranged from -9 to $+3$ dB in 3-dB steps. These noisy speech signals were denoted as "unprocessed." The processing was performed at a 16 kHz sampling rate, and the sounds were then upsampled to 48 kHz to ensure stable playback in a web-based experiment with/without an external digital-to-analog interface (See Sec. 3.3).

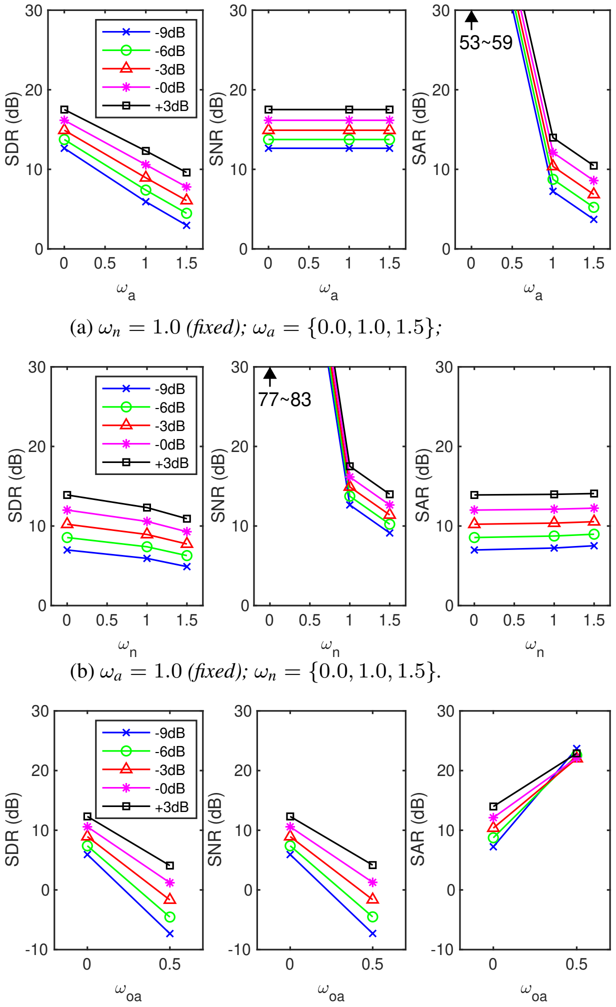
3.2. Speech enhancement front-end

SE was performed using the IRM (Eqs. (3) and (4)). The window length and the window shift for the STFT were 128 and 32 msec., respectively.

Next we created modified enhanced speech signals $\widehat{s}_\omega$ using Eqs. (9) and (10). Figure 2 shows the signal-to-distortion ratio (SDR), the signal-to-noise ratio (SNR), and the signal-to-artifact ratio (SAR) values [9] for modified enhanced signals $\widehat{s}_\omega$ when we controlled ω_{a} (Fig. 2a), ω_{n} (Fig. 2b), and ω_{oa} (Fig. 2c).

[2] Unlike Vincent et al. [9], this paper focuses on a 1-ch SE (noise reduction) task without interfering speakers, and thus reformulates the equations by focusing on only noise and artifact errors.

[3] (This footnote is for Fig. 2.) Even when $\omega_{\text{a}} = 0$ or $\omega_{\text{n}} = 0$, the SNRs and SARs were not ∞. This is presumably because SDR, SNR, and SAR were calculated after reconstructing modified signals $\widehat{s}_\omega$.



(a) $\omega_n = 1.0$ *(fixed);* $\omega_a = \{0.0, 1.0, 1.5\}$*;*

(b) $\omega_a = 1.0$ *(fixed);* $\omega_n = \{0.0, 1.0, 1.5\}$.

(c) *Observation adding*

Figure 2: *SDR, SNR, and SAR values [dB] for modified signals* $\hat{s}_\omega$*. Values outside the range are indicated by numerical values below the arrows, because SAR in (a) and SNR in (b) for* ω_a*=0 and* $\omega_n = 0$*, respectively, were quite large*[3]*. "x dB" in the legends denotes iSNR.*

3.3. Subjective test procedure

We conducted subjective tests using a crowdsourcing service provided by Lancers Co. Ltd. in Japan [22]. Any crowdworker could participate in the experimental task on a first-come-first-served basis. These experiments were performed using web pages developed for remote speech intelligibility experiments [23, 24, 11]. The participants were required to perform the experiments in a quiet place and to adjusted the volumes of their devices to an easily listenable level. They were instructed to write down the word that they heard using hiragana characters after listening the word once. There was a practice session for understanding the task, but no training session for advanced execution of the task.

We divided the subjective test into two sets: (i) an artifact-controlled set {unprocessed, $(\omega_\mathrm{n}, \omega_\mathrm{a})$=(1.0,1.0), (1.0,0.0), (1.0,1.5)} and (ii) a noise-controlled set {unprocessed, $(\omega_\mathrm{n}, \omega_\mathrm{a})$=(0.0,1.0), (1.5,1.0), observation adding}. The total number of presented stimuli was 400 words, comprising a combination of the four enhancement conditions of sets (i) or (ii), and the five iSNR $(-9, -6, -3, 0, +3$ dB) conditions with 20 words per condition. Each subject listened to a different word set, which was assigned randomly to avoid bias caused by word difficulty. For each set, 31 participants completed full tasks. Their native language was Japanese. None of them reported any hearing loss.

3.4. Speech recognition system

For a speech recognizer, we adopted ESPnet [25], which is open-source software for an end-to-end speech processing toolkit for building ASR systems that offer high performance against various benchmark tasks [26]. The ASR system was trained using training datasets comprised of the Corpus of Spontaneous Japanese (CSJ), one of the largest Japanese lecture speech corpora [27, 28]. For the training datasets, we used both clean signals and those contaminated with babble noise and impulse responses, which were recorded in the same rooms in Sec. 3.1. No enhanced signals were used for the ASR training. We trained the ASR system using the default setup of the ESPnet CSJ recipe [29], which exploits a state-of-the-art Transformer-based attention encoder-decoder model [30] with a connectionist temporal classification objective [31]. Since the speech of the evaluation data (FW07) is slower than the training data (CSJ), the values of the speed perturbation factor [32] were reduced to 0.8, 0.9, and 1.0 from their default values.

We used a default recurrent neural network language model trained according to the ESPnet CSJ recipe. For consistency with FW07 ' s evaluation data, we changed the output tokens of the ASR system from Japanese characters, which include both phonogram and ideograph characters, to *hiragana* characters, which are phonogram characters that roughly correspond to Japanese morae or consonant-vowel syllables.

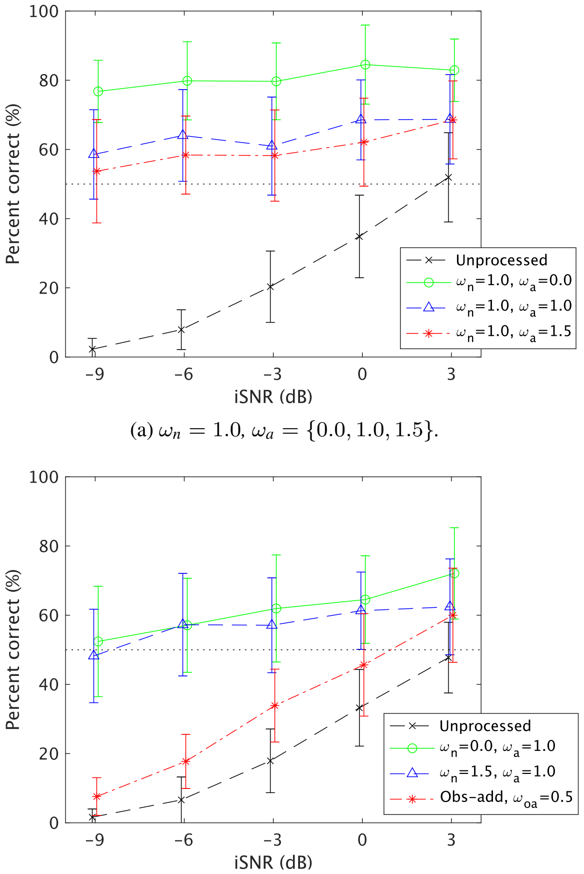
4. Results and discussions

4.1. Regarding artifacts and residual noise

Figures 3a and 3b show the results of the subjective tests when controlling ω_a and ω_n in Eq. (9), respectively. Comparing Figs. 3a and 3b, it can be seen that human intelligibility was highly affected when ω_a was controlled (Fig. 3a), on the other hand, the difference in intelligibility scores was not very large when $\omega_\mathrm{n} = 0.0$ and 1.5 (Fig. 3b). This tendency is similar to the results reported in [16], where the ASR performance was significantly affected by the artifact component and less affected by the noise error component.

Figure 4 compares the human and machine results. The subjective test results in Figs. 4a, 4c, and 4e are represented with a word *incorrect* rate (= 100 - correct rate, (%)) while the ASR results in Figs. 4b, 4d, and 4f are represented in word error rate (WER, (%)). The horizontal axes are the conditions of ω_a and ω_n in Eq. (9) and ω_oa in Eq. (10). Regarding the ASR performance (Figs. 4b and 4d), WER was highly degraded as ω_a increased; on the other hand, WER was only slightly affected by ω_n. This tendency is identical as in the subjective experiments shown in Figs. 4a and 4c. Similar results were previously reported [16].

It should be noted that artifact error $||\mathbf{e}_\mathrm{artif}||$ was larger than noise error $||\mathbf{e}_\mathrm{noise}||$, as shown from the SDRs for $\omega_a = 0$ and $\omega_n = 0$ in Fig. 2, and this may be one of the reasons why the artifacts had larger impacts on both humans and ASR. However, we still see that the artifacts had a significant impact on both the human and the ASR. The evaluation of the case where $||\mathbf{e}_\mathrm{noise}|| \simeq ||\mathbf{e}_\mathrm{artif}||$ while keeping the overall error amount (SDR) is one of our future works.



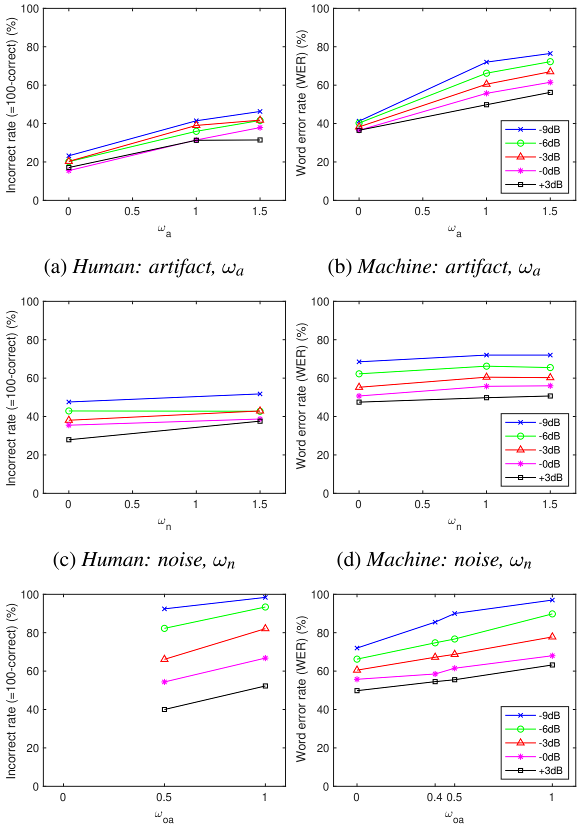
(a) $\omega_n = 1.0$, $\omega_a = \{0.0, 1.0, 1.5\}$.

(b) $\omega_a = 1.0$, $\omega_n = \{0.0, 1.5\}$*; observation adding.*

Figure 3: *Human subjective test results: mean and standard deviation of word correct rates (%).*

4.2. Regarding observation adding

The subjective test results for observation adding are shown as a red long dashed short dashed line in Fig. 3b. In our experiments, observation adding improved the intelligibility of the unprocessed signal (black dashed line); however, it highly degraded the intelligibility compared to the enhanced speech (green and blue lines). This is probably because the ω_{oa} (Eq. (10)) used here was 0.5, which is considerably larger than the practical SE post-processing (e.g., $\omega_{\text{oa}} < 0.1$). Actually, from Fig. 2c, the observation adding ($\omega_{\text{oa}} = 0.5$ in the figure) degraded SDR and SNR more than the original enhanced signal $\widehat{s}$ ($\omega_{\text{oa}} = 0$ in Fig. 2c), while it improved the SAR. That is, the impact of the degradation in SNR overrode the impact of the improvement in SAR; therefore, the intelligibility scores dropped. This result is also true for ASR in our experiments, as shown in Fig. 4f, that is, observation adding increased the WER. This result differs from that of Iwamoto et al. [16]. One of the reasons may be that our noise was recorded babble noise and Iwamoto et al. used public noise (from CHiME-3 [33]) [16], and the babble noise may make ASR more difficult. This suggests that further research is needed on the question of how to determine the preferable parameters of observation adding ω_{oa} for humans and machines (ASR) by looking at the type of noise and the SNR and SAR balance.



(a) *Human: artifact,* ω_a

(b) *Machine: artifact,* ω_a

(c) *Human: noise,* ω_n

(d) *Machine: noise,* ω_n

(e) *Human: Obs-add,* ω_{oa}

(f) *Machine: Obs-add,* ω_{oa}

Figure 4: *Human subjective test results (left column) and machine ASR results (WER [%]) (right column) as functions of* ω_a*,* ω_n*, and* ω_{oa}*. Note that vertical axis of human results is incorrect rate, i.e., 100 − correct rate (%), to make comparison with machine WER easier. "x dB" in the legends stands for the iSNR.*

5. Conclusions

We investigated the impact of residual noise and artifacts in enhanced speech signals on human intelligibility. It was shown that artifacts have a large impact on human intelligibility; on the other hand, the residual noise components have a lesser impact. This tendency is similar to the impact of SE on ASR. Our future work includes the development of SE approaches that are suitable both for humans and machines.

6. Acknowledgments

This research was partially supported by JSPS KAKENHI Nos. 21H03468 and 21K19794. The authors thank Mr. Kazuma Iwamoto and Mr. Shintaro Doan who respectively provided software for audio material creation and checked our subjective experimental results.

7. References

[1] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[2] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 577–581.

[3] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[4] M. Kolbaek, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[5] S. Xia, H. Li, and X. Zhang, "Using optimal ratio mask as training target for supervised speech separation," in *APSIPA ASC2017*, 2017, pp. 163–166.

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 deep noise suppression challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9271–9275.

[8] S. Venkataramani, R. Higa, and P. Smaragdis, "Performance based cost functions for end-to-end speech separation," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 350–355.

[9] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 14, no. 4, pp. 1462–1469, 2006.

[10] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[11] A. Yamamoto, T. Irino, S. Araki, K. Arai, A. Ogawa, K. Kinoshita, and T. Nakatani, "Effective data screening technique for crowdsourced speech intelligibility experiments: Evaluation with IRM-based speech enhancement," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2017 Asia-Pacific*, 2022, pp. 1402–1408.

[12] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 436–443.

[13] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," in *Interspeech*, 2018, pp. 1571–1575.

[14] T. Menne, R. Schlüter, and H. Ney, "Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6660–6664.

[15] M. Fujimoto and H. Kawai, "One-pass single-channel noisy speech recognition using a combination of noisy and enhanced features." in *Interspeech*, 2019, pp. 486–490.

[16] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR," in *Interspeech*, 2022, pp. 5418–5422.

[17] M. R. Schädler, A. Warzybok, S. D. Ewert, and B. Kollmeier, "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2708–2722, 2016.

[18] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech & Language*, vol. 48, pp. 51–66, 2018.

[19] K. Arai, A. Ogawa, S. Araki, K. Kinoshita, T. Nakatani, N. Kamo, and T. Irino, "Intelligibility prediction of enhanced speech using recognition accuracy of end-to-end asr systems," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1583–1589.

[20] M. Karbasi and D. Kolossa, "ASR-based speech intelligibility prediction: A review," *Hearing Research*, vol. 426, p. 108606, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378595522001745

[21] T. Kondo, S. Sakamoto, S. Amano, and Y. Suzuki, "Spoken word intelligibility tests under noisy conditions," in *40th International Congress and Exposition on Noise Control Engineering 2011, INTER-NOISE 2011*, 2011, pp. 1640–1646.

[22] "Lancers co. ltd." [Online]. Available: https://www.lancers.jp

[23] A. Yamamoto, T. Irino, K. Arai, S. Araki, A. Ogawa, K. Kinoshita, and T. Nakatani, "Comparison of remote experiments using crowdsourcing and laboratory experiments on speech intelligibility," in *Proc. Interspeech 2021*, 2021, pp. 181–185.

[24] T. Irino, H. Tamaru, and A. Yamamoto, "Speech intelligibility of simulated hearing loss sounds and its prediction using the Gammachirp Envelope Similarity Index (GESI)," in *Proc. Interspeech 2022*, 2022, pp. 3929–3933.

[25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *INTERSPEECH*, 2018, pp. 2207–2211.

[26] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on Transformer vs RNN in speech applications," in *ASRU*, 2019, pp. 449–456.

[27] S. Furui, K. Maekawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000, pp. 244–248.

[28] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[29] S. Watanabe, "ESPnet: End-to-end speech processing toolkit," https://github.com/espnet/espnet.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 6000—6 010.

[31] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving Transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in *INTERSPEECH*, 2019, pp. 1408–1412.

[32] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[33] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third ‘ CHiME ’ speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 504–511.