Multi-Metric Optimization Using Generative Adversarial Networks for Near-End Speech Intelligibility Enhancement

Haoyu Li, *Student Member, IEEE*, and Junichi Yamagishi, *Senior Member, IEEE*

***Abstract*—The intelligibility of speech severely degrades in the presence of environmental noise and reverberation. In this paper, we propose a novel deep learning based system for modifying the speech signal to increase its intelligibility under the equal-power constraint, i.e., signal power before and after modification must be the same. To achieve this, we use generative adversarial networks (GANs) to obtain time-frequency dependent amplification factors, which are then applied to the input raw speech to reallocate the speech energy. Instead of optimizing only a single, simple metric, we train a deep neural network (DNN) model to simultaneously optimize multiple advanced speech metrics, including both intelligibility- and quality-related ones, which results in notable improvements in performance and robustness. Our system can not only work in non-real-time mode for offline audio playback but also support practical real-time speech applications. Experimental results using both objective measurements and subjective listening tests indicate that the proposed system significantly outperforms state-of-the-art baseline systems under various noisy and reverberant listening conditions.**

***Index Terms*—Generative adversarial networks, multi-metric optimization, speech intelligibility.**

I. Introduction

REAL-LIFE speech communication, such as mobile telephony and public-address announcement, usually occurs in noisy and reverberant environments. These challenging environments severely degrade speech intelligibility, resulting in stressful listening or even non-understanding for listeners. Since noise sources are physically present in the near-end listener side, typical speech enhancement methods (e.g., [1]–[3]) which recover the clean speech from the noisy input, however, cannot be applied in such scenarios. As an alternative, there are many other methods aiming to modify the speech signal only to improve its intelligibility when exposed to noise and reverberation. In this paper, we refer to this task as near-end speech intelligibility enhancement.

Numerous algorithms for near-end intelligibility enhancement have been studied over the past decade (e.g., [4]–[10]). In particular, the 1st and 2nd Hurricane challenges [11], [12] summarized many effective algorithms and conducted comprehensive comparisons for each, providing remarkable reference value for researchers.

To increase intelligibility, many modification algorithms were designed on the basis of expert knowledge. For example, one algorithm called SSDRC [7] empirically sharpens the formant information and reduces the envelope variations of a speech signal, which leads to significant intelligibility improvement. Another example is a method called ASE [10], which maximizes intelligibility through certain audio manipulations, such as frequency-band decomposition and dynamic range compression, on the basis of sound engineering knowledge. Although these algorithms clearly improve speech intelligibility, they are dependent on domain experts' subjective experiences, thus still leaving room for improvement. These algorithms also consist only of non-parametric speech modifications; therefore, they cannot adapt well to the changing environments.

Inspired by human speech production characteristics, some algorithms (e.g., [13]–[15]) aim to convert normal speech to Lombard speech [16], which is naturally produced by speakers with increased vocal effort for higher intelligibility. To achieve speaking style conversion, most algorithms rely on vocoder-based analysis-and-synthesis techniques, where vocoder features are transformed to fit in the Lombard style. For example, Seshadri *et al.* [15] modified Mel-generalized cepstrum coefficients [17] of input speech to generate the Lombard-style speech by using log-domain pulse model vocoder [18]. However, using such a parametric vocoder inevitably degrades the converted speech quality. It was also found that even natural Lombard speech could only produce very limited intelligibility gains under low signal-to-noise ratio (SNR) conditions [11]. Consequently, the performance of such Lombard-inspired algorithms is still far from satisfactory.

Another group of algorithms were developed through optimizing certain objective intelligibility metrics. The basic concept is to modify the input speech in such a way as to maximize a target intelligibility metric under a known noise condition. For example, some algorithms (e.g., [5] and [8]) were proposed to maximize the speech intelligibility index (SII) [19]. Another

Manuscript received April 9, 2021; revised August 13, 2021; accepted September 5, 2021. Date of publication September 13, 2021; date of current version September 23, 2021. This work was supported in part by JST CREST VoicePersonae Project under Grant JPMJCR18A6, Japan, in part by MEXT KAKENHI under Grants 16H06302, 17H04687, 18H04120, 18H04112, 18KT0051, and 19K24372, Japan, and in part by SOKENDAI (The Graduate University for Advanced Studies), Japan. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding author: Haoyu Li.)*

The authors are with the National Institute of Informatics and with the Department of Informatics, SOKENDAI, Tokyo 101-8340, Japan (e-mail: haoyuli@nii.ac.jp; jyamagis@nii.ac.jp).

Digital Object Identifier 10.1109/TASLP.2021.3111566

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

group [6], [9], [20] optimizes a glimpse-based intelligibility metric [21]. These algorithms show promising results and do not rely on expert knowledge. Nevertheless, their performance still falls behind state-of-the-art algorithms such as SSDRC in subjective tests, as previously reported in [11]. This is because the objective metrics (e.g., SII) optimized within the above algorithms are relatively simple and inaccurate, i.e., they are not highly correlated with subjective intelligibility across different types of noise and other signal degradations [22]. Also, optimizing only a single target usually causes sub-optimality in another metrics, therefore limiting performance. Several advanced intelligibility metrics have recently been proposed and showed good results [23], [24]. However, it is still difficult to find closed-form solutions for optimizing these metrics due to mathematical complexities. Although numerical methods, such as gradient descent[1] and the genetic algorithm [25], can simultaneously optimize multiple complex metrics, their optimization schemes are based on offline iterative updates, and thus not suitable for real-time online applications.

Inspired by progresses in black-box function optimization [26], [27], we previously proposed a generative adversarial network (GAN)-based system [28] for near-end intelligibility enhancement. The system was composed of a generator that enhances the intelligibility of input speech and a discriminator that acts as a learned surrogate of evaluation metrics to guide the training scheme of the generator. We found that multiple intelligibility metrics could be effectively optimized with such a system. Nevertheless, this was still a non-causal system, and its generalization capability for unseen environments had not yet been extensively investigated.

In this paper, we propose a causal and light-weight system as an extension to our earlier system [28]. We substitute the original bidirectional long-short term memory (BLSTM) with causal convolution. To decrease the number of model parameters, we operate the speech signal on compact frequency-domain bands with equivalent rectangular bandwidth (ERB) scale resolution [29]. We use a GAN-based optimization scheme to jointly maximize not only multiple intelligibility metrics but also quality metrics for improved speech quality. We comprehensively evaluate the system's performance under different conditions with unseen noises and reverberations. Our experiments show that the improved system significantly increases the intelligibility and quality of speech over our original system [28] with far less parameters. Moreover, it also outperforms the state-of-the-art SSDRC baseline in both objective and subjective evaluations.

The rest of the paper is structured as follows. Section II describes the application scenario and formulates the problem of near-end speech intelligibility enhancement. Section III gives details on the proposed system. Section IV presents the experimental setup and results. We conclude this paper in Section V.

II. Scenario Description and Problem Formulation

Fig. 1 depicts an application scenario of near-end speech intelligibility enhancement. Let $x(n)$ be the input speech signal

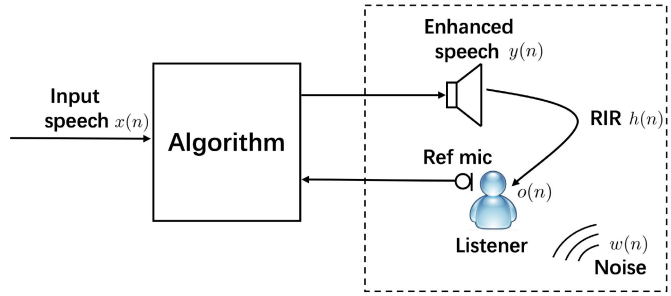


Fig. 1. Real-life scenario of near-end speech intelligibility enhancement.

with sampling index n. An algorithm is applied to modify $x(n)$, and then the enhanced signal $y(n)$ is output and played via a loudspeaker in a noisy and reverberant environment. The signal $o(n)$ observed by the near-end listener can thus be represented as

$$o(n) = y(n) * h(n) + w(n), \tag{1}$$

where $*$ denotes the convolution operation, $h(n)$ is the room impulse response (RIR),[2] and $w(n)$ is the additive noise disturbance. We further consider a common scenario in which the noise properties of $w(n)$ can be measured using a noise tracking algorithm via a reference microphone, such as the phone microphone for mobile telephony. On the other hand, we disregard the effect of reverberation, i.e., $h(n)$, since in practice it is difficult to estimate reverberation parameters in the presence of additive noise. With these assumptions, our target is thus to develop a system that transforms $x(n)$ into $y(n)$ to improve the intelligibility of $o(n)$ under a known noise condition.

More specifically, speech modification is carried out to redistribute the speech energy over time and frequency. Let $X(m,k)$ be the short-time Fourier transform (STFT) spectrogram of the raw signal $x(n)$, with the frame index m and frequency index k. We divide and group the frequency bins into the ERB-scaled bands [29] using triangular filter banks with the peak response being at the boundary between bands. Therefore, the input speech energy within one ERB band (indexed by band i at frame m) is given by

$$E_x(m,i) = \sum_k g_i(k)|X(m,k)|^2, \tag{2}$$

where $i \in \{1, 2, \ldots, I\}$ with I the total number of ERB-scaled bands, and $g_i(k)$ is the amplitude of the i-th triangular band at the k-th frequency bin. Similarly, we denote the spectrogram and energy band of noise as $W(m,k)$ and $E_w(m,i)$, respectively. The modified speech energy within one ERB band can be represented as $\alpha^2(m,i)E_x(m,i)$, where $\alpha(m,i)$ are the amplification factors that redistribute the speech energy across time and frequency bands. Due to the equal-power constraint, we also have the following equation with respect to $\alpha(m,i)$:

$$\sum_{m,i} \alpha^2(m,i)E_x(m,i) = \sum_{m,i} E_x(m,i). \tag{3}$$

[1]Gradient descent algorithm is inapplicable to the non-differentiable target metrics, where the gradients cannot be calculated.

[2]Loudspeaker response is integrated into the RIR for simplicity.

Next, the interpolated amplification factors applied to each frame m and frequency bin k are obtained by

$$\hat{\alpha}^2(m,k) = \sum_i g_i(k)\alpha^2(m,i). \tag{4}$$

They are then multiplied with the input spectrogram $X(m,k)$ to produce the enhanced spectrogram $\hat{\alpha}(m,k)X(m,k)$, which is subsequently converted to the enhanced signal $y(n)$ through the inverse STFT.

Instead of relying on expert knowledge to design an algorithm, we select several objective intelligibility and quality metrics as our optimization targets. We will further introduce the selected metrics in Section III-A. On the basis of the above discussion, we now reformulate the problem as follows. Given the noise estimation (in the form of $E_w(m,i)$) and the constraint of Equation (3), *our target is to find the amplification factors $\alpha(m,i)$ per time frame and ERB band to optimize the objective metrics of interest.*

III. Proposed System

In this section, we introduce our proposed GAN-based system to jointly optimize multiple speech metrics for improved intelligibility.

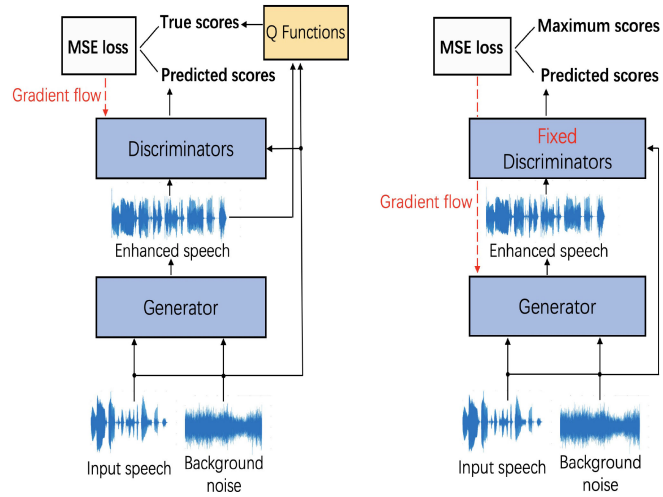
A. Target Speech Metrics

Objective metrics are used to measure the intelligibility of speech distorted by noise and reverberation. Very recently, Van Kuyk *et al.* [22] extensively tested the accuracy of many of these metrics by comparing their correlation coefficients with listening test scores. We accordingly selected the top three intelligibility metrics to build up and evaluate our proposed system. Their notations and brief descriptions are given as follows.

- *SIIB:* Speech intelligibility in bits (SIIB) [24] computes an estimation of the information shared between the clean and distorted speech signals in bits per second. Since an SIIB score relates to the signal duration, all stimuli are either repeated or truncated to have a consistent duration of 20 seconds when using SIIB, producing scores in the range of $[0, +\infty)$.
- *HASPI:* Hearing-aid speech perception index (HASPI) [30] estimates the intelligibility loss through the analysis of cepstral correlation and auditory coherence within an auditory model. To obtain the intelligibility score, we use a modified variant proposed in its recent improved version [23], where the final score, within the range of $[0, +\infty)$, is calculated as a weighted sum of the modulation filter outputs.
- *ESTOI:* Extended short-time objective intelligibility (ESTOI) [31] measures intelligibility by computing the correlation between the spectral of clean and distorted speech. ESTOI score ranges from 0 to 1.

In addition to these intelligibility metrics, we also selected the following two state-of-the-art quality metrics.

- *PESQ:* Perceptual evaluation of speech quality (PESQ) is a metric defined in ITU-T recommendation P.862 [32] for automated assessment of speech quality. PESQ score ranges from -0.5 to 4.5.



(a) Training process of discriminators (b) Training process of generator

Fig. 2. Diagram of the GAN model of the proposed system for near-end speech intelligibility enhancement.

- *ViSQOL:* Virtual speech quality objective listener (ViSQOL) is an objective metric released by Google [33], [34] for perceived audio quality. ViSQOL score ranges from 1 to 5.

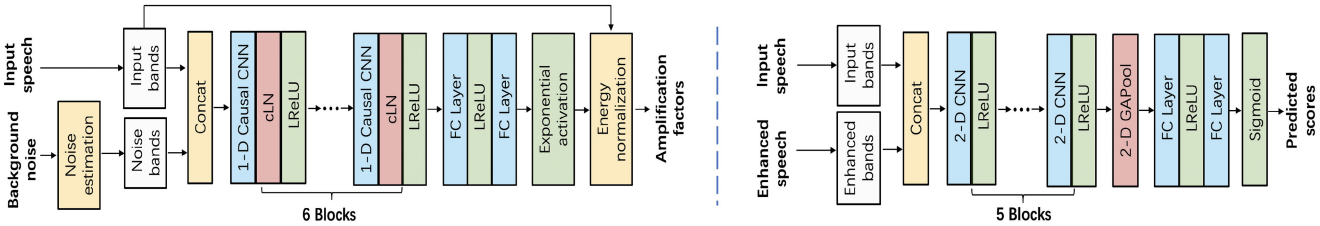
These two quality metrics are incorporated as the optimization targets to compensate for the quality loss caused by intelligibility-enhancing modifications.

All the above-mentioned metrics are so-called intrusive (or full-reference) models, which require a clean speech signal as the reference to predict an intelligibility or quality score for distorted speech. Although they can achieve high correlations with human subjective evaluations, they are too complex and mathematically intractable to handle. Particularly for a deep neural network (DNN) model, we cannot directly use such metrics as the training criteria since most of them are non-differentiable.[3] To overcome this obstruction, we thus introduce the following GAN model into our system.

B. System Overview

Fig. 2 shows the diagram of the GAN model of our proposed system. It is composed of a generator (G), an intelligibility discriminator (D_{int}), and a quality discriminator (D_{qua}). The G receives the input speech x and noise w and then outputs the enhanced speech $y = G(x,w)$, where we omit sampling index n from this point forward. Next, D_{int} and D_{qua} predict the intelligibility and quality scores of the enhanced speech, respectively. The predicted scores of the discriminators are expected to be close to the true scores calculated from the target objective metrics. Compared with those original metrics, which are quite complex, the gradients of DNN-based discriminators can be easily computed and back-propagated to G. Therefore, with the guidance of D_{int} and D_{qua}, G can be effectively trained to optimize the learned metrics of interest.

[3]ESTOI and PESQ metrics are technically differentiable under certain approximations, which have been studied in [35] and [36], respectively.



(a) G architecture (b) D_{qua} architecture

Fig. 3. Network architectures of our GAN model. Concat denotes concatenation operation. We set slope $= 0.3$ for all LReLU activations used in experiments.

More specifically, we now explain the training process of D_{int} in detail. As shown in Fig. 2(a), to predict the intelligibility scores, D_{int} takes three inputs: (1) the enhanced speech $G(x, w)$; (2) undistorted clean speech x; and (3) background noise w. We introduce the so-called Q functions to represent the target metrics to be modelled (described in Section III-A), with $Q_{\text{int}}(.)$ the functions for intelligibility metrics and $Q_{\text{qua}}(.)$ for quality metrics. Moreover, the signal example $\hat{y}$, which is pre-enhanced using other reference algorithms (e.g., SSDRC [7] and Opt-SII [5]), is also fed into D_{int} in the training. As demonstrated in our earlier study [28], learning such additional examples can stabilize the training process and improve performance. Given all the above notations, the loss function of D_{int} is represented as follows:

$$\mathcal{L}_D^{int} = \mathbb{E}_{x,w}\{[D_{\text{int}}(G(x,w),x,w) - Q_{\text{int}}(G(x,w),x,w)]^2 + [D_{\text{int}}(\hat{y},x,w) - Q_{\text{int}}(\hat{y},x,w)]^2\}. \quad (5)$$

By minimizing $\mathcal{L}_D^{int}$, D_{int} is encouraged to accurately predict the intelligibility scores. Similarly, we can represent the loss function of D_{qua} as Equation (6).

$$\mathcal{L}_D^{qua} = \mathbb{E}_{x,w}\{[D_{\text{qua}}(G(x,w),x) - Q_{\text{qua}}(G(x,w),x)]^2 + [D_{\text{qua}}(\hat{y},x) - Q_{\text{qua}}(\hat{y},x)]^2\} \quad (6)$$

Note that different from D_{int}, D_{qua} takes only two inputs: the enhanced speech $G(x, w)$ and clean reference speech x. This is because we have D_{qua} focus on measuring the quality of enhanced speech rather than the noisy observed speech.

Fig. 2(b) illustrates the G training process. We first fix the parameters of D_{int} and D_{qua}, and then apply the back-propagated gradients to update G to maximize the predicted intelligibility and quality scores. In order to increase the predicted scores as much as possible, we use the following loss function:

$$\mathcal{L}_G = \mathbb{E}_{x,w}\{[D_{\text{int}}(G(x,w),x,w) - t_{\text{int}}]^2 + \lambda[D_{\text{qua}}(G(x,w),x) - t_{\text{qua}}]^2\}, \quad (7)$$

where t_{int} and t_{qua} denote the maximum scores of the selected intelligibility and quality metrics, respectively, and λ is a hyper-parameter controlling the weight of speech quality to compensate for the quality degradation caused by intelligibility-enhancing modifications.

The generator (G) and discriminators (D_{int} and D_{qua}) are trained alternatively. At one training step, D_{int} and D_{qua} are trained individually with their corresponding loss functions, i.e., $\mathcal{L}_D^{int}$ and $\mathcal{L}_D^{qua}$. At the next training step, we fix the discriminators and only train G by minimizing loss $\mathcal{L}_G$. By this means, G can be effectively trained to optimize multiple advanced speech metrics, and the intelligibility of the enhanced speech (output by G) can be greatly improved and without too much quality degradation.

C. Network Architectures

The details of the network architectures are given in Fig. 3.

1) Generator: The input features for G are extracted from input speech and background noise. Specifically, the speech signal is transformed into the features containing 64 ERB-derived bands per time frame using Equation (2). There are two advantages for choosing features in the form of ERB-scaled bands rather than the raw frequency bins: (1) ERB-filterbank groups several perceptually-similar frequency bins into one band, producing a more robust feature;[4] and (2) The number of ERB bands is less than that of frequency bins, which can reduce the dimensions of the input and output features, resulting in a smaller model size. For background noise, we use the improved minima controlled recursive averaging algorithm (IMCRA) [37] to estimate noise power spectral density (PSD) $W^2(m, k)$, and then similarly extract 64 ERB bands as the noise features. These two features are then concatenated, resulting in a 128-channel feature vector and passed on to the following networks.

For network design, we choose the 1-D convolutional neural network (CNN) as the backbone for G due to the following reasons: (1) temporal convolution (1-D CNN with filter across time axis) has shown powerful modeling ability and been widely used in speech enhancement [38], [39] and synthesis [40]; and (2) the 1-D CNN is suited for real-time applications due to its low computational complexity.

As shown in Fig. 3(a), G consists of six blocks of causal 1-D CNN each with cumulative layer normalization (cLN) [38] and LeakyReLU activation (LReLU) [41]. The kernel size and output channels are set to (5, 256), (7, 256), (7, 256), (7, 256), (7, 256), and (5, 64), respectively. Two 64-node fully connected (FC) layers are subsequently followed by the last CNN block. The element-wise exponential activation function is then applied as follows:

$$output = \exp(3 * \tanh(u)), \quad (8)$$

where u is the result of the last FC layer, and the scale range of Equation (8) is approximately 0.05 to 20. The 64-dimensional

[4]Filterbank-based grouping operations are also implemented as front-end processing in many intelligibility metrics such as SIIB [24] and ESTOI [31].

output vector serves as the raw (non-normalized) amplification factors $\alpha(m, i)$, which redistribute the speech energy across time and frequency bands: the speech energy $E_x(m, i)$ (at frame m within band i) is boosted when $\alpha(m, i) > 1$; otherwise, suppressed. Furthermore, we add an energy normalization layer where the raw amplification factors are multiplied by a global scale factor γ in order to satisfy the equal-power constraint of Equation (3). Finally, the normalized $\alpha(m, i)$ are applied to reconstruct the enhanced speech signal, as described in Section II.

Except the last energy normalization operation, all layers in G are designed with causal configurations, which can run without dependencies of the future values of the signal. Moreover, G is a light-weight model containing only around 2.1 M parameters. It performs intelligibility enhancement very fast at the frame level, allowing for practical real-time speech applications. We will further discuss the extensions to real-time execution of our proposed system in Section IV-G.

2) Discriminators: Fig. 3(b) gives the detailed architecture of D_{qua}. It takes two types of ERB bands as input features: the unmodified input speech bands and enhanced bands. The D_{qua} is composed of five layers of 2-D CNN with the following kernel size and number of channels: [(1, 1), 8], [(3, 3), 16], [(5, 5), 32], [(7, 7), 48], and [(9, 9), 64], each with LReLU activation. A 2-D global average pooling (GAPool) [42] is added to the last CNN block to produce a fixed 64-dimensional output vector, which is then followed by an FC layer with 64 LReLU nodes. The last FC layer with sigmoid activation predicts the scores of modelled quality metrics, i.e., PESQ and ViSQOL. Thus, the number of nodes are accordingly set to 2. Similar to our previous study [28], we apply spectral normalization with 1-Lipschitz continuity [43] to all the layers used in D_{qua} to stabilize the training process.

For D_{int}, it shares the same network architecture with D_{qua}, except the inputs are changed to 3-channel features, i.e., (input, enhanced, noise), which requires an additional input of the estimated noise bands. Besides, the output nodes of D_{int} are set to 3, corresponding to the three intelligibility metrics to be modelled: SIIB, HASPI, and ESTOI.

IV. Experiments

A. Data Preparation

Speech materials consisted of Harvard sentences [44] spoken by two (one male [45] and one female [46]) native English speakers. The Harvard sentences are organized as 72 sets of 10 sentences each, and each set is designed to be phonetically balanced. Sentences were selected from sets 1–60, 61–66, and 67–72 for training, validation, and test data, respectively.

Six types of background noise were used: babble, restaurant, station, cafeteria, airport announcement, and speech-shaped noise (SSN), with the first five from the MS-SNSD dataset [47] and SSN artificially generated by VOICEBOX [48]. For training and validation data, we selected four types of noise (babble, station, restaurant, and SSN) to generate noisy speech at three SNR levels, i.e., –11, –7, and –3 dB. The remaining two types of noise were used for test data. For cafeteria noise, the SNRs were set to –9, –5, and –1 dB; for airport announcements noise, they were set to –13, –9, and –5 dB.

Although reverberation was disregarded in the training, we examined if the proposed system can work well in reverberant environments. Besides the original room condition (recorded in professional studios with reverberation time $T_{60} \approx 0.30$ s), another two RIRs were selected from a large room ($T_{60} = 0.61$ s) in the MIRD database [49] and stairway ($T_{60} = 0.92$ s) in the AIR database [50]. Thus, there were a total of three (1 original + 2 selected RIRs) reverberant environments considered in the test set. When generating noisy-reverberant speech, we first convolved the raw speech with the RIR, and then added the masker noise to the obtained reverberant speech at a desired SNR level.

To summarize, there were 14,400 (600 sentences $\times$ 2 genders $\times$ 3 SNRs $\times$ 4 noises) utterances in the training set; 1,440 (60 sentences $\times$ 2 genders $\times$ 3 SNRs $\times$ 4 noises) utterances in the validation set; and 2,160 (60 sentences $\times$ 2 genders $\times$ 3 SNRs $\times$ 2 noises $\times$ 3 reverberations) utterances in the test set. For the test set, a total of 18 listening conditions (comprising of 3 SNRs, 2 noises, and 3 reverberations) were extensively evaluated. It is worth noting that all the sentences, noises, reverberations (except the original condition), and SNR levels of the test set were unseen during model training.

B. Implementation Details

All signals were down-sampled to 16 kHz in our experiments. For feature extraction, we first used a Hanning window with a window size of 32 ms and hop size of 16 ms to compute the spectrogram. Next, 64 ERB-scaled triangular bands were applied to the spectrogram to produce the 64-dimensional input features for neural networks. All the input features were power-law compressed with a power of $1/6$. We chose SSDRC [7] as the reference algorithm to generate the signal example $\hat{y}$ that was used in Equations (5) and (6). During training, we normalized all metric scores to the range of $[0, 1]$, i.e., the same range with sigmoid activation, and set the target maximum scores (t_{int} and t_{qua} in Equation (7)) to 1. Specifically, we used the following parametric logistic function for score normalization:

$$f(v) = \frac{1}{1 + \exp(a * (v - b))}, \tag{9}$$

where v denotes the raw metric score. Parameters (a, b) were accordingly set as $(-0.06, 32)$ for SIIB; $(-0.95, 2.8)$ for HASPI; $(-8.0, 0.25)$ for ESTOI; $(-1.5, 2.5)$ for PESQ; and $(-2.5, 2.2)$ for ViSQOL. These parameters were empirically chosen to make the normalized scores uniformly distributed between 0 and 1, which helps reduce bias and stabilize GAN training.

For GAN model configurations, the Adam optimizer [51] was used in the training, with initial learning rates of 0.0004 and 0.0002 for the generator (G) and the discriminators (D_{int} and D_{qua}), respectively. The batch size was 1, and the hyper-parameter λ in Equation (7) was set to 0.5. The training process was terminated when all three intelligibility scores (SIIB, HASPI, and ESTOI) on the validation set stopped improving for five consecutive epochs.[5]

[5]Source codes and the pre-trained model are available at https://github.com/nii-yamagishilab/NELE-GAN

TABLE I
AVERAGE OBJECTIVE SCORES OF THE COMPARED SYSTEMS ACROSS DIFFERENT REVERBERANT CONDITIONS UNDER **CAFETERIA** NOISE

System	Intelligibility in $T_{60} \approx 0.30$ s				Intelligibility in $T_{60} = 0.61$ s				Intelligibility in $T_{60} = 0.92$ s				Quality	
	SIIB	HASPI	ESTOI	sEPSM	SIIB	HASPI	ESTOI	sEPSM	SIIB	HASPI	ESTOI	sEPSM	PESQ	ViSQOL
Unmodified	15.90	1.92	0.228	6.70	15.76	1.77	0.220	6.61	9.26	1.42	0.134	5.89	4.50	5.00
SSDRC	30.98	2.74	0.314	7.03	24.72	2.27	0.273	6.77	15.24	1.83	0.199	6.04	3.52	**2.71**
iMetricGAN	35.61	2.85	0.302	7.16	26.90	2.34	0.256	6.88	16.44	1.89	0.193	6.14	3.20	2.56
S-GAN	37.89	2.77	0.239	7.31	30.57	2.35	0.208	7.04	17.91	1.79	0.154	6.20	2.08	2.02
H-GAN	35.12	**3.12**	0.242	**7.55**	27.58	2.61	0.205	7.13	16.57	1.99	0.149	6.28	2.07	2.08
E-GAN	34.20	2.71	**0.331**	7.21	28.17	2.36	**0.285**	6.94	16.03	1.81	0.207	6.15	3.07	2.38
Proposed (S+H+E)	**41.33**	3.11	0.313	7.53	**32.99**	**2.62**	0.268	**7.17**	**18.90**	**2.00**	0.194	**6.28**	2.63	2.17
Proposed (All)	37.97	2.95	0.324	7.44	31.05	2.52	0.277	7.11	18.48	1.96	**0.209**	6.26	**3.54**	2.69

C. Objective Evaluations

In this section, we evaluated the proposed system through objective measurements. We first re-implemented several baseline systems, and then conducted an ablation test, yielding a total of eight systems evaluated in the experiments. We explain and notate each system as follows:

- *Unmodified:* Plain speech without any modification.
- *SSDRC:* A baseline system using the state-of-the-art SSDRC [7] algorithm, which achieved the highest and second highest intelligibility gains in the 1st [11] and 2nd [12] Hurricane challenges, respectively. It consists of two cascading non-parametric modifications: spectral shaping (SS) in frequency and dynamic range compression (DRC) in time.
- *iMetricGAN:* Our previously proposed system [28], in which we used BLSTM networks to optimize SIIB and ESTOI. We re-implemented it with the same model configurations, except the original sampling rate (44.1 kHz) was adjusted to 16 kHz. Its model size was 7.8 M parameters, which is much larger than the proposed system (2.1 M parameters for G).
- *S-GAN:* A system optimizing only SIIB, in which D_{int} was simplified to predict only a single SIIB score, and no D_{qua} was used for optimizing quality metrics.
- *H-GAN:* A system optimizing only HASPI.
- *E-GAN:* A system optimizing only ESTOI.
- *Proposed (S+H+E):* A partial version of our proposed system jointly optimizing three intelligibility metrics, i.e., SIIB, HASPI, and ESTOI. No D_{qua} was used for optimizing quality metrics.
- *Proposed (All):* Our full proposed system jointly optimizing three intelligibility metrics (SIIB, HASPI, and ESTOI) and two quality metrics (PESQ and ViSQOL).

We used the same target metrics (SIIB, HASPI, and ESTOI) as the evaluation measurements due to their very high correlations with human perception [22]. Moreover, we incorporated an additional advanced metric called sEPSM [52], i.e., an improved intelligibility prediction model based on the speech envelope-power spectrum model [53]. Note that sEPSM was completely unseen to the model, it was thus regarded as a third-party evaluation measurement in the experiments, producing scores in the range of $[0, +\infty)$. As discussed in Section IV-A, the objective intelligibility scores were extensively tested under two types of unseen noise under three room conditions: weak, medium, and severe reverberations.[6] The quality scores (PESQ and ViSQOL) were computed by comparing the enhanced speech (without noise and reverberation) with input unmodified speech. For the above-mentioned six measurements, higher scores indicate better performance.

Tables I and II list the average objective scores of each system under cafeteria and airport announcement noise, respectively. In both tables, **Proposed (All)** clearly outperformed the state-of-the-art baseline **SSDRC** in all room conditions with much higher intelligibility scores and comparable quality scores. Benefiting from new target metrics and network architecture, it also consistently improved upon the previously proposed **iMetricGAN** for all six measurements with a far smaller model size. Compared with **Proposed (S+H+E)**, **Proposed (All)** achieved much higher scores for speech quality with only a slight decrease in objective intelligibility scores.[7] **S-GAN**, **H-GAN**, and **E-GAN** performed well on their corresponding optimization targets. For example, we can see that **H-GAN** achieved the best HASPI scores in some cases. However, there still remains quite a bit of room for improvement in terms of other non-target metrics. This indicates that optimizing only a single metric might cause sub-optimality in those unconsidered metrics. By jointly optimizing multiple metrics, both **Proposed (S+H+E)** and **Proposed (All)** showed much more robust performance on all intelligibility measurements. Specifically, **Proposed (S+H+E)** produced the best results in terms of unseen sEPSM scores, and this further demonstrates that the multi-metric optimization strategy can lead to effective and generalized intelligibility improvement. More interestingly, we found that **Proposed (S+H+E)** and **Proposed (All)** can achieve extra SIIB gains even compared with the pure SIIB-oriented **S-GAN** system.

D. Subjective Listening Tests

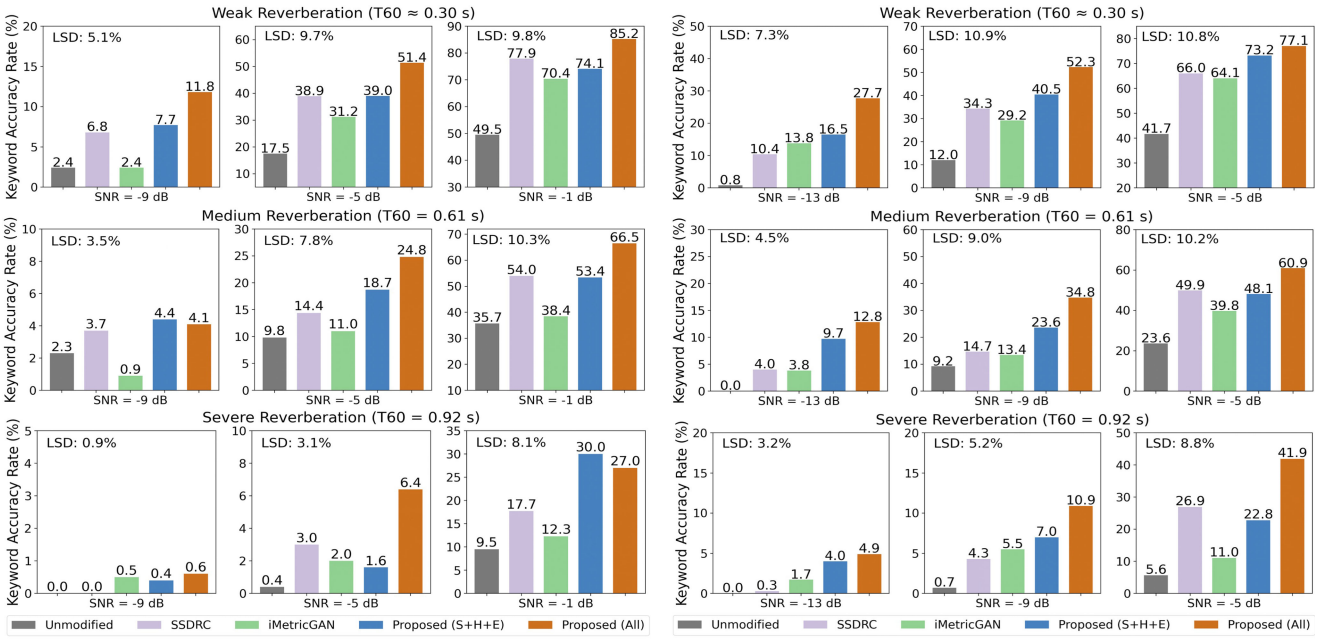
We conducted an intelligibility listening test to further evaluate the following five systems: **Unmodified**, **SSDRC**, **iMetricGAN**, **Proposed (S+H+E)**, and **Proposed (All)**.

[6] When computing the intelligibility scores under these reverberant conditions, the clean and distorted signals were time-aligned in advance.

[7] We also found that the quality scores can be further improved using a larger weight λ in Equation (7) at the cost of lower intelligibility scores.

TABLE II
AVERAGE OBJECTIVE SCORES OF THE COMPARED SYSTEMS ACROSS DIFFERENT REVERBERANT CONDITIONS UNDER **AIRPORT ANNOUNCEMENT** NOISE

System	Intelligibility in $T_{60} \approx 0.30$ s				Intelligibility in $T_{60} = 0.61$ s				Intelligibility in $T_{60} = 0.92$ s				Quality	
	SIIB	HASPI	ESTOI	sEPSM	SIIB	HASPI	ESTOI	sEPSM	SIIB	HASPI	ESTOI	sEPSM	PESQ	ViSQOL
Unmodified	16.25	2.20	0.191	6.63	16.12	2.07	0.190	6.61	9.43	1.58	0.115	5.79	4.50	5.00
SSDRC	32.49	3.38	0.286	7.24	25.80	2.71	0.261	6.85	16.37	2.17	0.203	6.06	3.52	**2.71**
iMetricGAN	35.68	3.44	0.280	7.37	27.72	2.73	0.250	6.95	17.98	2.23	0.204	6.18	3.22	2.58
S-GAN	42.34	3.54	0.214	7.82	34.21	2.85	0.195	7.26	21.75	2.25	0.160	6.30	2.12	2.04
H-GAN	39.19	**3.80**	0.226	7.89	31.50	3.03	0.201	7.34	20.25	**2.41**	0.165	6.37	2.08	2.10
E-GAN	35.04	3.36	0.283	7.39	28.88	2.82	**0.263**	7.03	18.09	2.23	0.205	6.17	3.07	2.40
Proposed (S+H+E)	**43.45**	3.75	0.279	**7.94**	**35.31**	**3.04**	0.250	**7.36**	**22.36**	2.40	0.206	**6.37**	2.71	2.19
Proposed (All)	42.54	3.72	**0.288**	7.87	34.30	3.00	0.257	7.30	22.03	2.38	**0.209**	6.36	**3.56**	2.67



(a) Under cafeteria noise.

(b) Under airport announcement noise.

Fig. 4. Mean keyword accuracy rates (KARs) in percentage points for each compared system across different listening conditions.

60 Harvard sentences (sets 67, 69, and 71 of the female speaker; and sets 68, 70, and 72 of the male speaker) were extracted and presented in each of the 18 listening conditions (3 SNRs $\times$ 2 noises $\times$ 3 reverberations), producing a total of 5,400 tested utterances (60 sentences $\times$ 18 conditions $\times$ 5 systems). We then divided these tested utterances into 90 blocks: each block consisted of 60 individual Harvard sentences, and each Harvard sentence was processed using a random system and under a random listening condition. A total of 90 native English speakers with no reported hearing impairments were recruited for the online test, and all were paid. Each participant was assigned to one block. They were instructed to listen to each tested utterance only once then type in as many words they heard as possible. We also implemented a cheater-detection mechanism by assigning five additional validation utterances (with very slight noise) to each block of the main listening test. Participants who did not reach 60% average word accuracy on these utterances were considered unqualified listeners, which led to three participants being excluded from the analysis. Following the evaluation rules of the 1st Hurricane challenge [11], we only accounted for the correct content words in each transcription by excluding the short common words: 'a,' 'the,' 'in,' 'to,' 'on,' 'is,' 'and,' 'of,' and 'for'. The keyword accuracy rate (KAR) was then computed as the performance measure of intelligibility.

The results are plotted in Fig. 4. Fisher's least significant difference (LSD) was also separately computed for each listening condition using ANOVAs to enable statistical comparisons of different systems. As shown in Fig. 4, modification algorithms can generally lead to substantial intelligibility gains to the unmodified speech, except for four extremely challenging conditions where all systems failed to reach 10% KAR. The best system in all but two of the 18 conditions was **Proposed (All)**. For all conditions, it consistently outperformed not only **iMetricGAN**, but also the state-of-the-art **SSDRC**. Interestingly, although its objective intelligibility scores were lower than those of **Proposed (S+H+E)** (see in Tables I and II), it showed

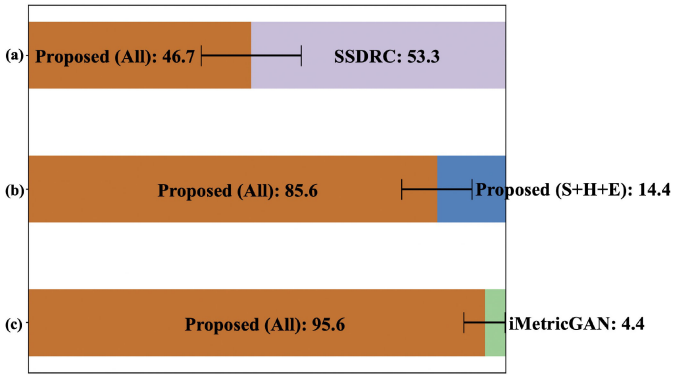


Fig. 5. Preference scores (%) with 95% confidence intervals on speech quality compared between **Proposed (All)** and three reference systems.

much higher increases in KAR. This reveals that incorporating quality metrics into training can largely contribute to subjective intelligibility, which is likely due to the effective suppression of audible artefacts.[8]

We also conducted AB preference tests to evaluate the perceptual quality of the enhanced speech. We conducted pairwise comparisons between **Proposed (All)** and the following three systems: (1) **SSDRC**; (2) **iMetricGAN**; and (3) **Proposed (S+H+E)**. 90 enhanced samples were randomly selected from the test set for each system, and a total of 15 listeners participated. Each participant was instructed to listen to 18 randomized sample pairs, and for each pair they had to select the one that sounded better in terms of speech quality. As we can see from Fig. 5, **Proposed (All)** achieved significantly higher preference scores than **iMetricGAN** and **Proposed (S+H+E)** and performed comparably with **SSDRC**. Such results clearly indicate that speech quality can be effectively improved through incorporating objective quality metrics into model training.

E. Acoustic Analysis on Enhanced Speech

We analyzed the acoustic properties of the enhanced speech. For deeper insight, we used **SSDRC** as the reference system to conduct a comparative study. Fig. 6 gives examples of waveforms and spectrograms for different signals. From the spectrograms, we found that both **SSDRC** and **Proposed (All)** modified the speech signal through redistributing its energy from low frequencies to the middle and high frequencies. By comparing Fig. 6(c) with (b), **Proposed (All)** tended to allocate more energy on the middle-frequency regions (2∼4 kHz) of the voiced segments (see black dashed box), while **SSDRC** emphasized the high-frequency regions (4∼8 kHz) of the unvoiced segments (see blue dashed box). We can also see that the waveform envelope of the enhanced speech from **Proposed (All)** is similar to that of the original unmodified speech. In contrast, the modified waveform of **SSDRC** drastically changed, resulting in more acoustic artefacts.

[8]Audio samples of the tested systems are available at https://nii-yamagishilab.github.io/hyli666-demos/intelligibility/index.html

TABLE III
AVERAGE OBJECTIVE SCORES ON NEW GERMAN SPEAKER TEST SET

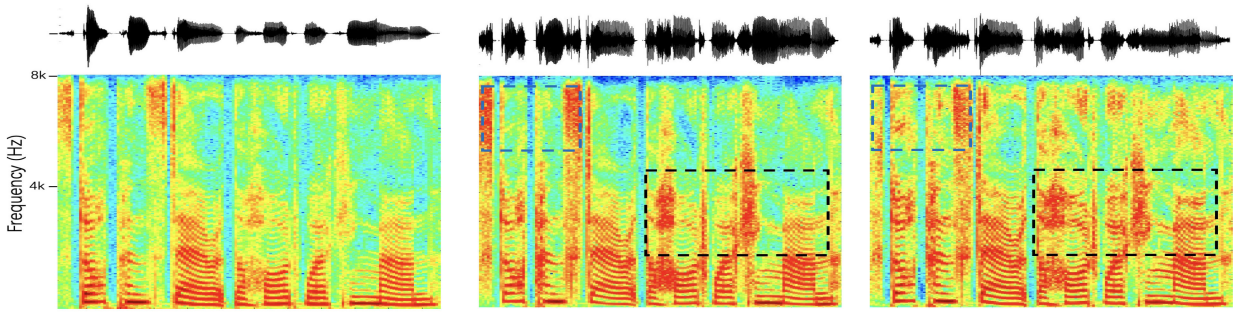
System	Intelligibility				Quality	
	SIIB	HASPI	ESTOI	sEPSM	PESQ	ViSQOL
Unmodified	12.64	1.63	0.167	6.65	4.50	5.00
SSDRC	25.27	2.40	0.252	7.00	3.40	2.58
Proposed (All)	**28.94**	**2.66**	**0.254**	**7.44**	**3.46**	**2.81**

We also investigated the gain (in dB) of the long-term average spectrum (LTAS) calculated over one unmodified utterance. The gain values indicate the energy level of a signal in a certain frequency region: the signal energy is higher than the unmodified utterance with gain > 0 dB; otherwise, lower. As shown in Fig. 7, frequency regions from 1 kHz to 8 kHz were effectively boosted in both **SSDRC** and **Proposed (All)**, which accords with our observations in Fig. 6. Different from noise-independent **SSDRC**, **Proposed (All)** can adapt well to the changing environments. For example, the noise in Fig. 7(b) was extremely strong (up to 40 dB gain) in the low-frequency regions (∼125 Hz). Thus, the system automatically gave up much more speech components in these regions, compared with how it performed under weaker noise in Fig. 7(a). We also found that the speech components between 65 Hz to 150 Hz were particularly boosted under cafeteria noise, as shown in Fig. 7(a). Interestingly, this coincides with the properties of the cafeteria noise where a peak gain was also exhibited near the same regions (see blue and black lines). We hypothesize that by increasing the speech components in such narrow but noise-dominant regions, the target speech can be differentiated from the surrounding noise in an easier manner through achieving a certain perception threshold. On the other hand, **SSDRC** performed merely the same processing of the speech with two different noises (see red lines); therefore, it cannot make full use of additional noise information. This is one of the points explaining why our proposed system performed better in both objective and subjective evaluations.

F. Analysis of System Robustness

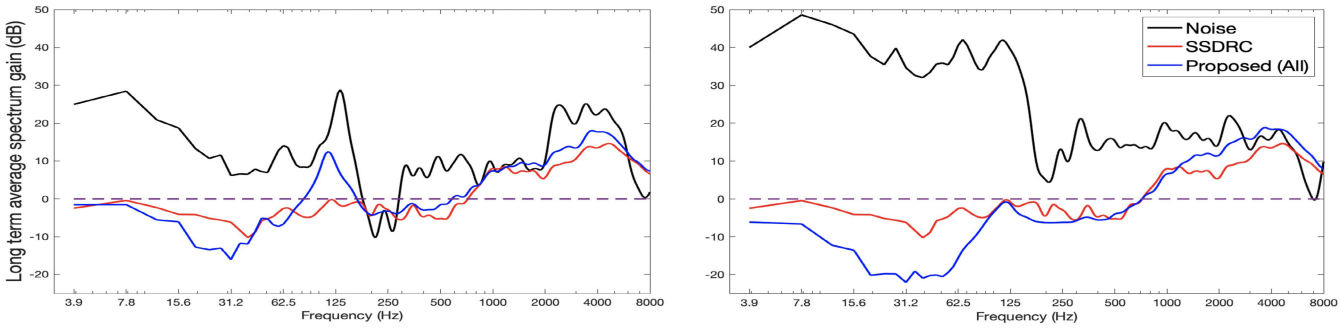
We further analyzed the system's robustness in two particular situations, where (1) speaker and language are unseen to the model; and (2) background noise estimation is not accurate.

1) Speaker and Language Generalization: We tested the proposed system on a separately-created German speaker test set to examine if it can work under the mismatched speaker and language conditions. Specifically, we extracted 100 clean utterances from an unseen male German speaker [12] and set the same 18 listening conditions (i.e., 2 noise types, 3 SNRs and 3 room conditions) as used in the original test set (see Section IV-A), resulting in a total of 1,800 tested utterances. Table III lists the objective evaluation results on this new German speaker test set, where the scores were averaged over all listening conditions. We can see that even though **Proposed (All)** was built only upon English training data, it still achieved



(a) Unmodified (b) SSDRC (c) Proposed (All)

Fig. 6. Waveforms and their spectrograms on one utterance under cafeteria noise at SNR=–5 dB for different signals: (a) Unmodified input speech, (b) enhanced speech from **SSDRC**, and (c) enhanced speech from **Proposed (All)**. Utterance used is notated as "f_70_8," i.e., the 8-th utterance in 70-th list of female speaker.



(a) LTAS gain under cafeteria noise at SNR=–5 dB (b) LTAS gain under airport announcement noise at SNR=–13 dB

Fig. 7. Long-term average spectrum (LTAS) gain (dB) over LTAS of unmodified utterance (f_70_8) for: (1) masker noise, (2) enhanced speech from **SSDRC**, and (3) enhanced speech from **Proposed (All)**.

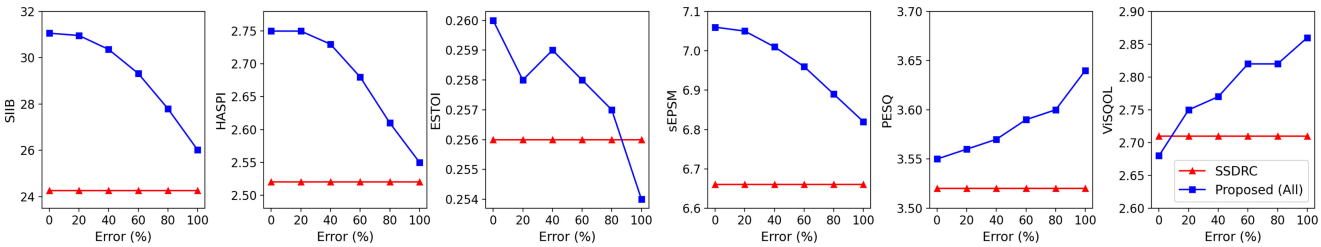


Fig. 8. Average objective scores as noise estimation error is artificially added to noise PSD.

significant intelligibility gains and outperformed **SSDRC** by a large margin. This further demonstrates that the proposed system is robust, which can generalize well to mismatched speaker and language.

2) Tolerance to Noise Estimation Error: Next, we measured the tolerance of the proposed system to inaccuracy of background noise estimation. As described in Section II, in order to exploit noise information, our system requires a reference microphone and runs IMCRA [37] algorithm to estimate noise PSD, i.e., $W^2(m,k)$. However, such noise estimation might be inaccurate, for example, when the noise is highly non-stationary or the reference microphone is distant from the listener's position. To simulate estimation error in this process, we randomly marked certain noise PSD bins as error bins with an error rate of $\epsilon\%$; thus, the corrupted noise PSD $W_e^2(m,k)$ is given as follows:

$$W_e^2(m,k) = \begin{cases} exp(N), & if\ \ error \\ W^2(m,k), & else \end{cases} \tag{10}$$

where N is the random noise generated from Gaussian distribution with the same mean and variance as those of $logW^2(m,k)$, and error rate controls the corruption level: a higher $\epsilon\%$ indicates that each estimated bin is more likely filled with random noise, making noise estimation more inaccurate.

Fig. 8 shows the objective metric scores under different error rates. For the intelligibility metrics (i.e., SIIB, HASPI, ESTOI, and sEPSM), the corrupted noise PSD did not affect performance much when the error rate ϵ was less than 40% but decreased

intelligibility scores incrementally when $\epsilon > 40\%$. However, even when noise estimation completely failed (i.e., $\epsilon = 100\%$), **Proposed (All)** could still surpass the performance of **SSDRC** in intelligibility metrics (except ESTOI), which demonstrates that the proposed system is very robust against noise estimation error. From another point of view, by simply substituting random values for noise PSD, **Proposed (All)** degenerates into a noise-independent system. This also indicates that our system is flexible and can adapt to scenarios in which the implementation of a reference microphone is not available. More interestingly, we found that the quality metrics (PESQ and ViSQOL) increased with increasing noise estimation error. We hypothesize that the system tends to modify the speech in a relatively aggressive manner to fully make use of noise information, e.g., giving up much more speech components in low-frequency regions when low-frequency noise is strong (see blue line in Fig. 7(b)). For larger $\epsilon\%$, the system cannot exploit useful information as the given noise PSD becomes random; therefore, it tends to perform moderate modification, resulting in higher quality scores.

G. Extensions to Real-Time Execution

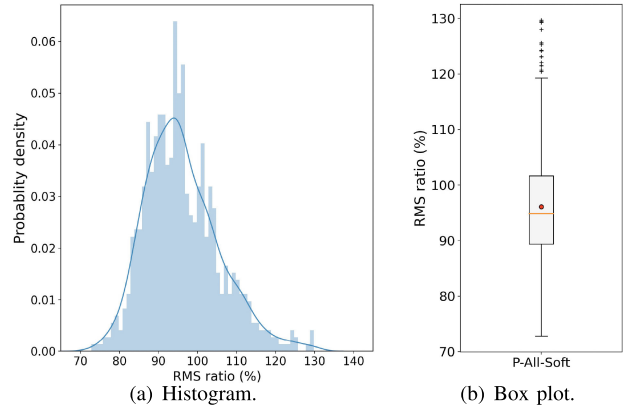
Real-time execution is crucial for many speech applications such as mobile telephony. In this section, we discuss the causality of the proposed system in detail. As discussed in Section III-C1, the G used in **Proposed (All)** can inherently perform intelligibility enhancement at the frame level in a causal manner. However, due to the equal-power constraint of Equation (3), we still need to collect the entire signal to calculate the global energy of an utterance. Thus, we consider two extended methods to overcome this limitation.

First, we revised the original utterance-level normalization (Equation (3)) to the following frame-level normalization:

$$\sum_i \alpha^2(m,i)E_x(m,i) = \sum_i E_x(m,i), \quad \forall m. \tag{11}$$

As shown in Equation (11), the energy is normalized at each frame m instead of the whole utterance, which enables the system to perform real-time execution under the equal-power constraint. We denote this modified frame-level normalization method for our proposed system as **P-All-FL**. Compared with the original proposed system with the utterance-level normalization method (denoted as **P-All-UL**), **P-All-FL** can only redistribute the speech energy across the frequency bands within one frame but not perform inter-frame redistribution.

Second, we consider another normalization method for application scenarios in which the equal-power constraint is not rigorous. As mentioned in Section III-C1, the global scale factor γ originally used in **P-All-UL** is calculated dynamically for each utterance to achieve perfect energy normalization. With this method, however, we prepare such a γ in advance by statically calculating the average energy ratio between the unmodified and enhanced speech over the whole training set. The γ is then applied to the raw amplification factors to compensate for the energy loss, achieving a soft energy normalization where the enhanced speech has approximately the same energy with the unmodified one. We denote this method as **P-All-Soft**, and



(a) Histogram. (b) Box plot.

Fig. 9. Statistical results (γ used in **P-All-Soft** was set to 5.62) of RMS ratios between enhanced and unmodified raw speech: (a) frequency density histogram of RMS ratios and (b) box plot on RMS ratios, with red dot representing mean score.

γ was calculated as 5.62 from the training set. Fig. 9 presents the statistical results of the root-mean-square (RMS) ratios between the enhanced and unmodified speech on the test set. As shown in Fig. 9, the distribution of RMS ratios was concentrated close to one with a very small deviation. This indicates that the energy of enhanced speech can be well maintained within the approximately same level as the unmodified one by using **P-All-Soft** method.

Table IV lists the objective evaluation results on the three normalization methods. The scores were averaged over the whole test set across three SNR levels, three room conditions, and two unseen noises. We found that **P-All-FL** did provide intelligibility gains to the unmodified speech. However, it performed much worse than the other two methods due to the lack of inter-frame energy distribution, which further reveals that energy reallocation in time is crucial for intelligibility improvement. Although **P-All-Soft** cannot perfectly fulfill the equal-power constraint, it satisfies the causality requirement and showed a comparable performance to **P-All-UL**. Note that all three methods differed only in the energy normalization strategy, while the core model of G used in the experiments was identical. By choosing a suitable normalization method in accordance with actual needs, the proposed system can satisfy different requirements of causality and energy constraint.

Finally, we give a brief analysis on the system complexity. The enhancement module, i.e., G, is composed of 2.1 M weight parameters. Since each weight is used once for one multiply-add operation per frame (16 ms), G thus takes 262.5 million floating-point operations per second (MFLOPS) for real-time execution.[9] For other main modules, including two FFTs (for input speech and background noise analyses, respectively), one inverse FFT (for enhanced speech reconstruction), and IMCRA noise estimation, they take around 4.0 MFLOPS. The total complexity of the proposed system is around 270 MFLOPS. Considering both model size and the computational complexity, our proposed system is light-weight and can be easily implemented in practice.

[9]One multiply-add operation is counted as two operations.

TABLE IV
AVERAGE OBJECTIVE SCORES FOR SYSTEMS WITH DIFFERENT NORMALIZATION METHODS ON TEST SET

Normalization method	Causal	Equal-power constrained	Intelligibility				Quality	
			SIIB	HASPI	ESTOI	sEPSM	PESQ	ViSQOL
Unmodified	–	–	13.79	1.82	0.180	6.37	4.50	5.00
P-All-UL	×	✓	31.06	2.75	0.260	7.06	3.55	2.68
P-All-FL	✓	✓	20.15	2.26	0.193	6.73	3.29	2.53
P-All-Soft	✓	×	29.79	2.68	0.249	7.06	3.55	2.67

V. CONCLUSION

We proposed a GAN-based system for near-end speech intelligibility enhancement. To generate the intelligible and high-quality speech, we introduced a GAN model into our system to jointly optimize multiple intelligibility and quality metrics. Three modules are used in the GAN model to carry out such multi-metric optimization: an intelligibility discriminator that learns to predict the objective intelligibility scores of speech as accurately as possible, quality discriminator that similarly learns to predict the objective quality scores, and a generator that enhances the input speech signal to maximize both intelligibility and quality scores, which are computed with the above two discriminators, respectively.

Experimental results from both objective measurements and large-scale listening tests indicated that the proposed system can lead to significant intelligibility gains and perform much better than several compared baselines. It also generalizes well to various listening environments including unseen noises and reverberations. Moreover, the system is light-weight with only 2.1 M parameters and can be easily extended to enable real-time execution.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[4] K. Nathwani, M. Daniel, G. Richard, B. David, and V. Roussarie, "Formant shifting for speech intelligibility improvement in car noise environment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5375–5379.

[5] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[6] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech*, 2012, pp. 955–958.

[7] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, 2012, pp. 635–638.

[8] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time SII," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 5, pp. 851–862, May 2015.

[9] Y. Tang and M. Cooke, "Learning static spectral weightings for speech intelligibility enhancement in noise," *Comput. Speech Lang.*, vol. 49, pp. 1–16, 2018.

[10] C. Chermaz and S. King, "A sound engineering approach to near end listening enhancement," in *Proc. Interspeech*, 2020, pp. 1356–1360.

[11] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: The hurricane challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.

[12] J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-enhancing speech modifications - the hurricane challenge 2.0," in *Proc. Interspeech*, 2020, pp. 1341–1345.

[13] K. Nathwani, G. Richard, B. David, P. Prablanc, and V. Roussarie, "Speech intelligibility improvement in car noise environment by voice transformation," *Speech Commun.*, vol. 91, pp. 17–27, 2017.

[14] A. R. López, S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Speaking style conversion from normal to lombard speech using a glottal vocoder and Bayesian GMMs," in *Proc. Interspeech*, 2017, pp. 1363–1367.

[15] S. Seshadri, L. Juvela, P. Alku, and O. Räsänen, "Augmented CycleGANs for continuous scale normal-to-lombard speaking style conversion," in *Proc. Interspeech*, 2019, pp. 2838–2842.

[16] H. Lane and B. Tranel, "The lombard sign and the role of hearing in speech," *J. Speech Hear. Res.*, vol. 14, no. 4, pp. 677–709, 1971.

[17] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - A unified approach to speech spectral estimation," in *Proc. Int. Conf. Spoken Lang. Process.*, 1994, pp. 1043–1046.

[18] G. Degottex, P. Lanchantin, and M. Gales, "A log domain pulse model for parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 57–70, 2017.

[19] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.

[20] C. Valentini-Botinhao, J. Yamagishi, S. King, and R. Maia, "Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 665–686, 2014.

[21] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoustical Soc. America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[22] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2153–2166, Nov. 2018.

[23] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI) Version 2," *Speech Commun.*, vol. 131, pp. 35–46, 2021.

[24] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.

[25] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992.

[26] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. Int. Conf. Mach. Learn. PMLR*, 2019, pp. 2031–2041.

[27] M. Kawanaka, Y. Koizumi, R. Miyazaki, and K. Yatabe, "Stable training of dnn for speech enhancement based on perceptually-motivated black-box cost function," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7524–7528.

[28] H. Li, S.-W. Fu, Y. Tsao, and J. Yamagishi, "iMetricGAN: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning," in *Proc. Interspeech*, 2020, pp. 1336–1340.

[29] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1/2, pp. 103–138, 1990.

[30] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, 2014.
[31] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
[32] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, 2001, vol. 2, pp. 749–752.
[33] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, and A. Hines, "Objective assessment of perceptual audio quality using ViSQOLAudio," *IEEE Trans. Broadcast.*, vol. 63, no. 4, pp. 693–705, Dec. 2017.
[34] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *Proc. IEEE 12th Int. Conf. Qual. Multimedia Experience*, 2020, pp. 1–6.
[35] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.
[36] J. Kim, M. El-Kharmy, and J. Lee, "End-to-end multi-task denoising for joint SDR and PESQ optimization," 2019, *arXiv:1901.09146*.
[37] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[38] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[39] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6875–6879.
[40] Y. Ren *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2020, *arXiv:2006.04558*.
[41] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.
[42] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.
[43] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.
[44] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.
[45] C. Valentini-Botinhao, C. Mayo, and M. Cooke, "Hurricane natural speech corpus - higher quality version," 2019. [Online]. Available: https://doi.org/10.7488/ds/2482
[46] P. Demonte, "HARVARD speech corpus - Audio recording 2019," 2019. [Online]. Available: https://doi.org/10.17866/rd.salford.c.4437578.v1
[47] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," in *Proc. Interspeech*, 2019, pp. 1816–1820.
[48] M. Brookes, "Voicebox: Speech processing toolbox for MATLAB," 2005. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[49] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 313–317.
[50] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. Digit. Signal Process.*, 2009, pp. 1–5.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[52] K. Steinmetzger, J. Zaar, H. Relaño-Iborra, S. Rosen, and T. Dau, "Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations," *J. Acoustical Soc. Amer.*, vol. 146, no. 4, pp. 2562–2576, 2019.
[53] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoustical Soc. Amer.*, vol. 134, no. 1, pp. 436–446, 2013.

Haoyu Li (Student Member, IEEE) received the B.Sc. degree in physics from Nanjing University, Nanjing, China, in 2016 and the M.Eng. degree in electronic engineering from The University of Tokyo, Tokyo, Japan, in 2018. He is currently working toward the Ph.D. degree with the SOKENDAI/National Institute of Informatics, Japan. His current research interests include statistical machine learning and speech enhancement.

Junichi Yamagishi (Senior Member, IEEE) received the Ph.D. degree from the Tokyo Institute of Technology (Tokyo Tech), Tokyo, Japan, in 2006. During 2007–2013, he was a Research Fellow with the Centre for Speech Technology Research, University of Edinburgh, Edinburgh, U.K. In 2013, he was appointed as an Associate Professor with the National Institute of Informatics (NII), Japan. He is currently a Professor with NII. He is also a Honorary Professor with the University of Edinburgh. Since 2000, he has authored and coauthored more than 300 refereed papers in international journals and conferences. He was awarded the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, and the Young Scientists' Prize from the Minister of Education, Science and Technology, the JSPS prize, the Docomo Mobile Science Award in 2010, 2013, 2014, 2016, and 2018. He served previously as the co-organizer for the bi-annual ASVspoof challenge and the bi-annual Voice conversion challenge. He was also a Member of the IEEE Speech and Language Technical Committee during 2013–2019, an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING during 2014–2017, and a Chairperson of ISCA SynSIG during 2017–2021. He is currently a PI of JST-CREST and ANR supported VoicePersona project and a Senior Area Editor of the IEEE/ACM TASLP.