Speech Enhancement-assisted Voice Conversion in Noisy Environments

Yun-Ju Chan*, Chiang-Jen Peng*, Syu-Siang Wang†, Hsin-Min Wang‡, Yu Tsao§ and Tai-Shih Chi*
* Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
† Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan
E-mail: sypdbhee@saturn.yzu.edu.tw
‡ Institute of Information Science, Academia Sinica, Taipei, Taiwan
§ Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
E-mail: yu.tsao@citi.sinica.edu.tw

***Abstract*—Numerous voice conversion (VC) techniques have been proposed for the conversion of voices among different speakers. Although good quality of the converted speech can be observed when VC is applied in a clean environment, the quality degrades drastically when the system is run in noisy conditions. In order to address this issue, we propose a novel speech enhancement (SE)-assisted VC system that utilizes the SE techniques for signal pre-processing, where the VC and SE components are optimized in an joint training strategy with the aim to provide high-quality converted speech signals. We adopt a popular model, StarGAN, as the VC component and thus call the combined system as EStarGAN. We test the proposed EStarGAN system using a Mandarin speech corpus. The experimental results first verified the effectiveness of joint training strategy used in EStarGAN. Moreover, EStarGAN demonstrated performance robustness in various unseen noisy environments. The subjective listening test results further showed that EStarGAN can improve the sound quality of speech signals converted from noise-corrupted source utterances.**

I. Introduction

The goal of voice conversion (VC) is to convert a human voice from a source speaker into that of a predefined target, but preserve the linguistic content of the source speech [1], [2]. By applying a conversion function in terms of the speaker attributes to achieve a specific goal, numerous VC approaches have been proposed in the past, such as Gaussian mixture model (GMM) based methods [1], [3] and exemplar-based techniques [4], [5]. Several studies leverage deep-learning-based techniques to promote VC to conduct non-linear transformations and synthesize high-quality speaker voices, which are much similar to the target voice [6]–[9]. For example, generative adversarial network (GAN) based approaches, such as CycleGAN [10] and StarGAN [11], apply a generator composed of an encoder and a decoder to perform VC, while using an additional discriminator or classifier to handle the statistical distribution of the target speaker.

In general, the quality of the converted speech is reasonable in a clean background; however, the quality of VC applied in a noisy environment degrades dramatically. To address this issue, several works trained their systems directly under an interference background during the training stage; hence, robust VC could be achieved in online testing [12]–[15]. For such VC systems with denoising function, a noisy source feature is placed at the input of the system; the target clean voice is then obtained at the output.

Instead of deriving the VC model directly from the noisy–clean training pairs, we propose a novel approach that first enhances the source features through a speech enhancement (SE) model and then converts the enhanced features into the target speaker's voice. SE technology is an essential speech processing front-end that extracts clean components from the noise-corrupted input for downstream applications, such as hearing aids [16], [17], automatic speech recognition [18], [19], and speaker recognition [20]–[22]. Several SE approaches have also been incorporated with deep-learning-based neural network architectures to improve the system performance in non-stationary noisy environments [23]–[25]. By applying a model that considers the contextual information from the input logarithmic power spectra (LPS), decent speech quality and intelligibility can be obtained [18].

We proposed a novel E-StarGAN VC system that integrates an SE function to handle a noisy background during VC tasks. The LSTM-based model structure was used to construct the SE function, whereas the following StarGAN architecture was used to conduct the conversion. AutoVC, which has been proven to provide a good VC performance in a clean environment, is used as the generator in the StarGAN. The experiment results clearly indicate that the newly proposed E-StarGAN VC model significantly improves the sound quality and similarity with the target speaker of the converted speech. In addition, E-StarGAN has proven to be a robust VC system under various noisy environments. In the future, we plan to study E-StarGAN for singing voice conversion. We investigated EStarGAN in terms of objective and subjective tests under different noisy conditions. The evaluation results suggest that when compared with the baseline AutoVC, the converted speech signals yield by EStarGAN, which jointly trained the SE and VC models, have lower Mel cepstral distortion (MCD) values and high signal quality and similarity. Moreover, by further optimizing the discriminator and

978-616-590-477-3 ©2022 APSIPA

classifier, we can obtain even better conversion performance.

II. Related work

The conventional StarGAN-VC model architecture is composed of generator $\mathbb{G}$, discriminator $\mathbb{D}$, and classifier $\mathbb{C}$ modules. The VC process is carried out through the generator, which is composed of an encoder and a decoder. The encoder is used to decompose a source acoustic feature $\mathbf{y}$ into a speaker-independent vector; it is then combined with the target speaker attribute $\mathbf{I}_t$ for the decoder to reconstruct the target feature $\mathbf{t}$. Notably, $\mathbf{I}_t$ is a one-hot vector, where the nonzero element corresponds to the target speaker identity. Therefore, we can formulate the VC process as $\mathbf{t} = \mathbb{G}\{\mathbf{y}, \mathbf{I}_t\}$. The discriminator and classifier functions in the training stage are used as additional regularizers to perform StarGAN-VC. Specifically, adversarial, identity-mapping, speaker classification, and cycle-consistency losses [11] are used and introduced in the following subsections.

A. Adversarial loss function

With respect to the discriminator and generator, the adversarial loss function is defined as

$$\begin{aligned} \mathcal{L}^{\mathbb{D}}_{adv} &= -E_{\mathbf{t},\mathbf{I}_t}\{log[\mathbb{D}(\mathbf{t},\mathbf{I}_t)]\} - E_{\mathbf{y},\mathbf{I}_t}\{log[1-\mathbb{D}(\mathbb{G}(\mathbf{y},\mathbf{I}_t),\mathbf{I}_t)]\}, \\ \mathcal{L}^{\mathbb{G}}_{adv} &= -E_{\mathbf{y},\mathbf{I}_t}\{log[\mathbb{D}(\mathbb{G}(\mathbf{y},\mathbf{I}_t),\mathbf{I}_t)]\}, \end{aligned} \tag{1}$$

where E represents the expectation operation. We then minimize $\mathcal{L}^{\mathbb{D}}_{adv}$ for $\mathbb{D}$ to distinguish the real speech $\mathbf{t}$ from a fake sample $\mathbb{G}(\mathbf{y})$, and minimize $\mathcal{L}^{\mathbb{G}}_{adv}$ for $\mathbb{G}$ to generate a speech feature in the target domain that can deceive the discriminator. The discriminator output is a two-dimensional one-hot vector that indicates the true or false class of the associated input.

B. Speaker-classification loss function

To further preserve the speaker characteristics from the generated speaker voice, classifier $\mathbb{C}$, which specifies the speaker identity from K speakers is applied to the output side of the generator. Thus, the output dimension of $\mathbb{C}$ is K. The corresponding loss function is formulated as follows:

$$\begin{aligned} \mathcal{L}^{\mathbb{C}}_{cls} &= -E_{\mathbf{t},\mathbf{I}_t}\{log[p_{\mathbb{C}}(\mathbb{C}(\mathbf{t}))]\}, \\ \mathcal{L}^{\mathbb{G}}_{cls} &= -E_{\mathbf{y},\mathbf{I}_t}\{log[p_{\mathbb{C}}(\mathbb{C}(\mathbb{G}(\mathbf{y},\mathbf{I}_t)))]\}. \end{aligned} \tag{2}$$

The loss is then minimized for the classifier to identify the input speaker and for the generator $\mathbb{G}$ to provide a voice that is similar to the target speaker.

C. Cycle-consistency loss function

In addition to applying the adversarial loss for StarGAN, the cycle-consistency loss was leveraged for a generator to preserve more acoustic content during the conversion process. The cycle-consistency loss function is defined as follows:

$$\mathcal{L}^{\mathbb{G}}_{cyc} = E_{\mathbf{y},\mathbf{I}_t,\mathbf{I}_s}\{\|\mathbb{G}(\mathbb{G}(\mathbf{y},\mathbf{I}_t),\mathbf{I}_s) - \mathbf{y}\|_1\}, \tag{3}$$

where $\mathbf{I}_s$ represents the source speaker attribute. Thus, the generated output speech feature from the generator belongs to the source domain according to the provided source speaker information, $\mathbf{I}_s$.

D. Identity-mapping loss function

To further promote the generator for learning speech structures, the identity-mapping loss is used and defined in the following equation:

$$\mathcal{L}^{\mathbb{G}}_{idm} = E_{\mathbf{y},\mathbf{I}_s}\{\|\mathbb{G}(\mathbf{y},\mathbf{I}_s) - \mathbf{y}\|_1\}. \tag{4}$$

From (4), we expect the output of the generator $\mathbb{G}(\mathbf{y},\mathbf{I}_s)$ to be identical to that of the model input $\mathbf{y}$.

III. The Proposed Approach

In this study, we proposed EStarGAN for achieving VC in noisy environments. Similar to StarGAN, EStarGAN consists of a generator, discriminator, and speaker classifier. The detailed EStarGAN model is shown in Fig. 1, wherein the generator architecture is illustrated in the upper row, and the discriminator and speaker classifier are depicted at the bottom left and right, respectively. As mentioned in the Sec. II, the discriminator was applied to determine whether the input was true speech samples or the generated fake one, whereas the speaker classifier was employed to conduct the speaker identification task. Therefore, we focused on introducing the generator and training procedure used in this subsection.

A. Generator

From Fig. 1, the generator $\mathbb{F}$ is composed of SE and AutoVC components, where AutoVC comprises the encoder, decoder, and post-net processor. In addition, we formulated the VC process as $\hat{\mathbf{t}} = \mathbb{F}(\mathbf{x}, \mathbf{I}_s, \mathbf{I}_t)$ with respect to the noisy input, $\mathbf{x}$.

1) Feature enhancement: The BLSTM model was used to perform SE, which is composed of two BLSTM-hidden layers and followed by the 80-node feed-forward layer. For the noise-suppression task, we passed the noisy LMS $\mathbf{x}$ through the FE to provide an enhanced feature $\hat{\mathbf{y}}$ from the output of the SE model. The mean square error (MSE) $\mathcal{L}_{mse}\{\hat{\mathbf{y}}, \mathbf{y}\}$ criterion was then applied to minimize the distance between $\hat{\mathbf{y}}$ and $\mathbf{y}$ for applying the BLSTM model during the training stage. In this study, 160 cells were used for each BLSTM hidden layer.

2) Auto-encoder based voice conversion: The block diagram of AutoVC is depicted in Fig. 1(b)–(d) and the conversion process is formulated as $\mathbb{F}_{auto}$. As the figure indicates, the encoder, $\mathbb{F}_{\mathbb{E}}$, encoded the enhanced LMS $\hat{\mathbf{y}}$ and the associated source speaker attribute $\mathbf{I}_s$ to generate the speaker-independent embedding feature, $\mathbf{c} = \mathbb{F}_{\mathbb{E}}(\hat{\mathbf{y}}, \mathbf{I}_s)$. This encoded feature comprised the outputs of the forward and backward paths in the latest BLSTM hidden layer. We then applied a 16-factor down-sampling followed by an up-sampling operation on the embedding feature along the time axis and then concatenated its outcome with the target speaker attribute $\mathbf{I}_t$ for the following decoder. The decoder output was the converted target LMS $\tilde{\mathbf{t}}$; this was filtered in the post-net thereafter to provide $\hat{\mathbf{t}}$ in the final output to shrink the possible signal distortions. Consequently, the relationship between input $\hat{\mathbf{y}}$ and output $\hat{\mathbf{t}}$ is $\hat{\mathbf{t}}, \tilde{\mathbf{t}} = \mathbb{F}_{auto}(\hat{\mathbf{y}}, \mathbf{I}_s, \mathbf{I}_t)$. In

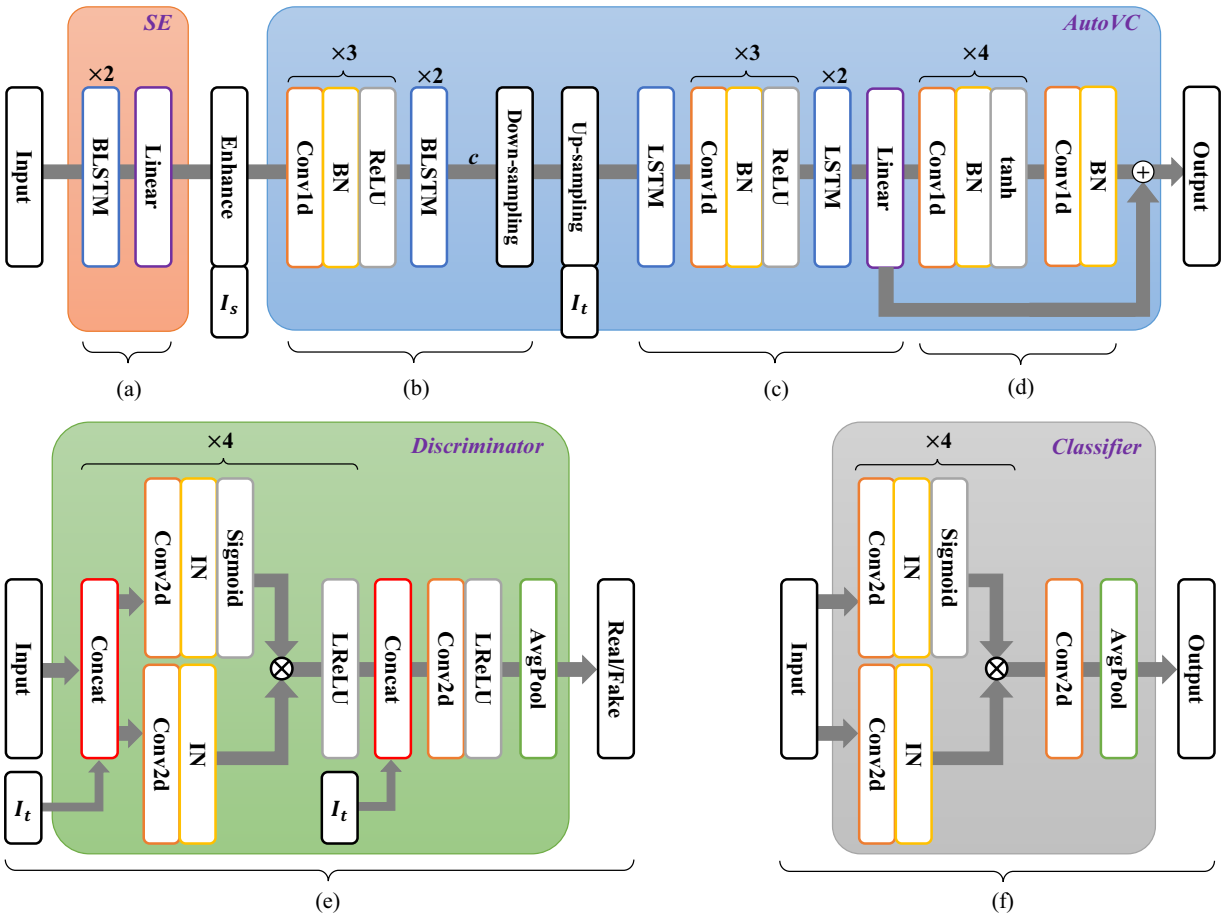


Fig. 1. Block diagram of the proposed EStarGAN VC, which comprises six sub-systems: (a) SE, (b) encoder, (c) decoder, (d) post-net, (e) discriminator, and (f) speaker classifier. "IN" and "BN" represents the instance and batch normalization techniques, respectively [26]. Notably, the upper row of the figure represents a block diagram of the generator (SE and AutoVC), wherein the input is a source noise feature, and the output is converted clean one.

addition to $\hat{\mathbf{t}}$, $\hat{\dagger}, \tilde{\dagger} = \mathbb{F}_{auto}(\hat{\mathbf{y}}, \mathbf{I}_s, \mathbf{I}_s)$ was also used to conduct a loss function that was defined as:

$$\mathcal{L}_{auto} = \mathcal{L}_{mse}\{\tilde{\dagger}, \hat{\mathbf{y}}\} + \mathcal{L}_{mse}\{\hat{\dagger}, \hat{\mathbf{y}}\} + \lambda_{auto}\mathcal{L}_{mse}\{\mathbb{F}_{\mathbb{E}}(\hat{\dagger}, \mathbf{I}_s), \mathbf{c}\}. \quad (5)$$

Notably, $\mathbb{F}_{\mathbb{E}}(\hat{\dagger}, \mathbf{I}_s)$ in (5) represents the output $\hat{\dagger}$ was placed in the input side of $\mathbb{F}_{\mathbb{E}}$ to generate the associated speaker-independent embedding feature.

B. Training stage

1) Loss functions: The loss functions, listed in (6) were leveraged and minimized to perform EStarGAN VC.

$$\begin{aligned} \mathcal{L}^{\mathbb{F}} &= \mathcal{L}^{\mathbb{F}}_{adv} + \lambda_{cls}\mathcal{L}^{\mathbb{F}}_{cls} + \lambda_{cyc}\mathcal{L}^{\mathbb{F}}_{cyc} + \lambda_{idm}\mathcal{L}^{\mathbb{F}}_{idm}, \\ \mathcal{L}^{\mathbb{C}} &= \mathcal{L}^{\mathbb{C}}_{cls}, \\ \mathcal{L}^{\mathbb{D}} &= \mathcal{L}^{\mathbb{D}}_{adv} + \lambda_{gp}\mathcal{L}^{\mathbb{D}}_{gp}, \end{aligned} \quad (6)$$

where

$$\begin{aligned} \mathcal{L}^{\mathbb{D}}_{adv} &= -E_{\mathbf{t},\mathbf{I}_t}\{log[\mathbb{D}(\mathbf{t}, \mathbf{I}_t)]\} - E_{\mathbf{y},\mathbf{I}_t}\{log[1 - \mathbb{D}(\mathbb{F}(\mathbf{x}, \mathbf{I}_s, \mathbf{I}_t), \mathbf{I}_t)]\}, \\ \mathcal{L}^{\mathbb{F}}_{adv} &= -E_{\mathbf{x},\mathbf{I}_s,\mathbf{I}_t}\{log[\mathbb{D}(\mathbb{F}(\mathbf{x}, \mathbf{I}_s, \mathbf{I}_t), \mathbf{I}_t)]\}, \\ \mathcal{L}^{\mathbb{F}}_{cls} &= -E_{\mathbf{x},\mathbf{I}_s,\mathbf{I}_t}\{log[p_{\mathbb{C}}(\mathbb{C}(\mathbb{F}(\mathbf{x}, \mathbf{I}_s, \mathbf{I}_t)))]\}, \\ \mathcal{L}^{\mathbb{F}}_{cyc} &= E_{\mathbf{x},\mathbf{I}_s,\mathbf{I}_t}\{\|\mathbb{F}_{auto}(\mathbb{F}(\mathbf{x}, \mathbf{I}_s, \mathbf{I}_t), \mathbf{I}_t, \mathbf{I}_s) - \mathbf{y}\|_1\}, \\ \mathcal{L}^{\mathbb{F}}_{idm} &= E_{\mathbf{y},\mathbf{I}_s}\{\|\mathbb{F}(\mathbf{x}, \mathbf{I}_s, \mathbf{I}_s) - \mathbf{y}\|_1\}. \end{aligned} \quad (7)$$

In addition, $\mathcal{L}^{\mathbb{D}}_{gp}$ proposed in [27] was used as an additional regulizer for the discriminator in EStarGAN.

2) Training steps: The detailed training process consists of four steps: (1) We first extracted the LMS features for each clean and noise utterance. (2) By placing the noisy LMS features at the input, we pre-train the SE module for $220k$ iterations in order to generate the corresponding clean features. (3) Next, we cascade the SE module and Auto-VC (encoder, decoder and post-net) module in Fig. 1 and jointly train the two models based on Eq. (5) for $905k$ training steps. (4) Finally, we fixed the SE module and train the discriminator, classifier, and Auto-VC iteratively based on Eq. (6) for $380k$ epochs.

Notably, if we only implement the training steps (1), (2), and (3), we can also obtain a noise-robust VC system, termed "*jt*-SE+VC". In this case, noisy LMS features in the online stage are processed by the concatenated and jointly trained SE+VC to generate the converted speech signals. For comparison, we perform a direct concatenation of SE and VC, which are trained individually and without joint-training. More specifically, the SE is trained using noisy-clean paired

utterances, and the VC is trained using paired source-target clean utterances. Then, we directly cascade the SE and VC modules; hereafter, this direct concatenation setup is denoted as "SE+VC".

IV. Experiments and Analysis

In the following subsections, we first introduce the experimental setup and then provide results along with discussions.

A. Experiment setup

The Taiwan Mandarin Hearing in Noise Test [28] script, which contains 320 sentences was used to form our database. The training and testing scripts were pronounced by four male and four female Mandarin speakers ($K = 8$ in Sec. II-B) and recorded at a 16 kHz sampling rate in a quiet meeting space to provide 2,560 utterances. We selected 310 samples per speaker and prepared 2,480 clean utterances to provide the training set among these recordings. Each training waveform was artificially deteriorated by 100 different types of noises [29] signal-to-noise ratios (SNRs) of 15, 10, and 5 dB to finally provide 744,000 noisy–clean pairs for model training. Meanwhile, the other 80 clean recordings were contaminated by "engine," "pink," "street," and "white" noises, which selected from [30] and freesound[1], at three SNRs of 5, 10, and 15 dB to generate the corresponded noisy testing data. As a result, 960 noisy utterances were used for testing.

We applied a short-time Fourier transformation with a 64-ms frame size and a 16-ms hop length to decompose a waveform for the feature extraction process. The prepared spectral features were then filtered using mel filters and generated 80-dimensional LMS for this study. For a waveform synthesis from the LMS, we used the parallel WaveGAN model proposed in [31]. The model was pre-trained using the COSPRO corpus [32] and then retrained on the training database described in this study. The detailed EStarGAN model architecture was illustrated in Fig. 1. For the one-dimensional convolutional layer (Conv1d) used in the generator of EStarGAN, the channel, kernel, and stride sizes were 512, 5, and 1, respectively. The rest configuration was introduced as follows:

- **Encoder**: The cell size of each of two LSTM hidden layers was 512.
- **Decoder**: The 512, 1024, and 1024 cells were sequentially used to construct hidden LSTM layers. The output layer was an 80-node feed-forward network.
- **Post-net**: The latest output feed-forward layer contained 80 nodes.
- **Discriminator**: Five two-dimensional convolutional (Conv2d) layers were stacked in order of [32, (3,9), (1,1)], [32, (3,8), (1,2)], [32, (3,8), (1,2)], [32, (3,6), (1,2)] and [40, (80,5), (1,1)], which follow the setting representation "[channel size, kernel size, stride size]". The dimension of the final output from the applied average pooling (AvgPool) process was one.

[1]https://freesound.org/

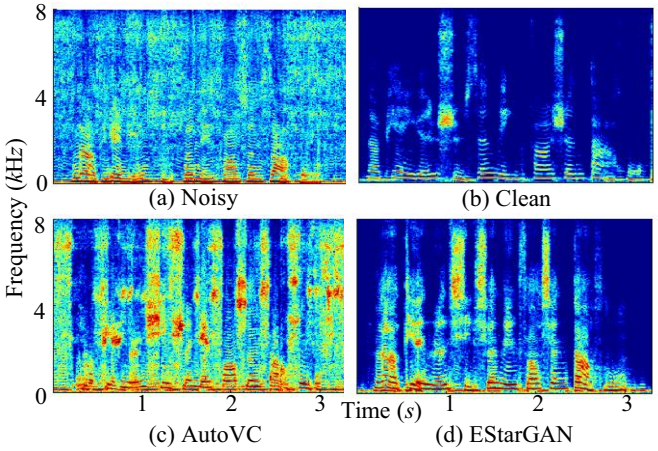


Fig. 2. The spectrograms of (a) Source noisy, (b) Target clean and the source noisy speech converted by (c) AutoVC and (d) EStarGAN.

- **Speaker classifier**: Five Conv2d layers were used in order of [8, (2,2), (4,4)], [16, (2,2), (4,4)], [32, (2,2), (4,4)], [16, (1,2), (5,4)] and [8, (1,2), (6,4)]. The AvgPool operation finally provided an eight-dimensional output.

The model configuration of AutoVC was identical to that used in the EStarGAN generator. Notably, despite the prepared parallel database, the training process for AutoVC and EStarGAN was conducted under non-parallel conditions. Finally, we applied the MCD objective metric to evaluate the proposed system.

B. Experiment results

To visualize VC performances, we present the LPS of a single source noisy utterance in Fig. 2(a), as well as the clean target speech in Fig. 2(b). The noisy utterances converted by AutoVC and EStarGAN are depicted in Figs. 2(c) and (d), respectively. In addition, the noisy source and clean target voices were pronounced by two different speakers. Comparing Figs. 2(a)(b)(c), the AutoVC spectrogram shows a clear acoustic structure in the voice parts, while introducing additional noise components in the silence segments. By contrast, EStarGAN converted speech in Fig. 2(d) shows a cleaner LPS and more detailed harmonic structures than those in 2(c).

We evaluated the conversion performances of (a) male to male (M2M), (b) male to female (M2F), (c) female to male (F2M), and (d) female to female (F2F) in terms of the MCD

TABLE I
AVERAGED MCD VALUES OF AUTOVC, SE+VC, *jt*-SE+VC AND ESTARGAN IN M2M, F2F, M2F AND F2M CONDITIONS.

	M2M	F2F	M2F	F2M	Avg.
AutoVC	9.43	9.70	10.03	9.48	9.66
SE+VC	9.45	9.71	9.75	9.56	9.62
***jt*-SE+VC**	7.74	7.95	8.17	8.35	8.05
EStarGAN	**7.48**	**7.83**	**7.92**	**8.17**	**7.85**

metric, as shown in Table. I. For this comparison, the average MCD values between the EStarGAN-converted source speech and the target ground truth were denoted as "EStarGAN". The scores between the AutoVC-converted voices and the target speech were represented as "AutoVC". In addition, we listed the MCD results of SE+VC and *jt*-SE+VC in the table for more comparisons. The average result among all testing utterances for each VC system is also listed in the table (denoted as "Avg.").

From the table, SE+VC (directly concatenating SE and VC modules without joint-training) slightly improve the performance of AutoVC in terms of the Avg value. This observation suggests that the enhanced source feature cannot bring clear benefits for the following AutoVC-based StarGAN. One possible explanation is that the statistical properties of an enhanced speech features do not match well with the following VC generator because of the independent training processes for SE and VC. Contrarily, both *jt*-SE+VC and EStarGAN consistently yield superior results to the baseline AutoVC and SE+VC under all testing conditions, as well as for Avg. The results demonstrate that the advantages of the joint-training strategy. Moreover, EStarGAN outperforms *jt*-SE+VC,suggesting that further improvement can be attained by performing joint-training on the discriminator and classifier in addition to the AutoVC generator.

Next, we evaluate AutoVC, SE+VC, *jt*-SE+VC and EStarGAN in terms of the average MCD values under four noisy environments. The results were showed in Table II. Notably, the tested noisy conditions differed from those in the model training stage. From the table, EStarGAN provides the lowest MCD values for all testing noisy conditions when compared to other VC approaches. The result confirms the effectiveness of the joint training strategy and demonstrates the robustness of the proposed EStarGAN in performing noisy speech conversion.

TABLE II
AVERAGED MCD VALUES OF AUTOVC, SE+VC, *jt*-SE+VC AND ESTARGAN IN ENGINE, PINK, WHITE, STREET NOISE ENVIRONMENTS.

	engine	**pink**	**white**	**street**
AutoVC	9.55	9.66	9.79	9.63
SE+VC	9.61	9.62	9.63	9.62
***jt*-SE+VC**	8.03	8.05	8.08	8.05
EStarGAN	**7.83**	**7.84**	**7.88**	**7.86**

TABLE III
AVERAGED MOS SCORES OF AUTOVC, SE+VC, *jt*-SE+VC AND ESTARGAN IN WHITE, STREET NOISE ENVIRONMENTS.

	white	**street**	**Avg.**
AutoVC	1.91	1.85	1.88
EStarGAN	**3.66**	**3.64**	**3.65**

TABLE IV
AVERAGED SIM VALUES OF AUTOVC, SE+VC, *jt*-SE+VC AND ESTARGAN IN WHITE, STREET NOISE ENVIRONMENTS.

	white	**street**	**Avg.**
AutoVC	2.63	2.85	2.73
EStarGAN	**1.94**	**1.98**	**1.96**

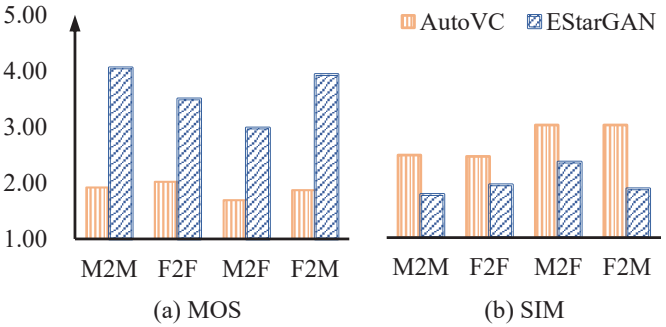


Fig. 3. The (a) MOS score and (b) SIM value for AutoVC and EStarGAN during subjective tests.

In addition to the objective evaluations, we also conducted subjective listening tests in terms of the mean opinion score (MOS) and similarity (SIM) for AutoVC and EStarGAN. The value range of each MOS and SIM was $[1, 5]$. A higher MOS score indicates a better sound quality. However, a lower SIM value leads to a higher similarity. A single-blind listening test was conducted on 15 untrained but experienced normal hearing subjects who did not know which VC system was used. Two audio clips comprising a converted speech and a reference sound were presented for each participant. They were then asked to give an MOS score for the converted speech and the SIM value to determine the level of similarity for the heard samples. Listening samples were selected from the testing database and prepared in the following processes. The source utterances pronounced by two male and two female speakers were contaminated by white and street noises at an SNR of 10 dB and then converted into the clean target speech, which was uttered by all eight speakers. From these voice samples, we selected 40 EStarGAN and 40 AutoVC converted voices to conduct this subjective test. The averaged MOS and SIM with respect to the four gender-based conditions are illustrated in Fig. 3. The results of MOS and SIM with respect to both noises are listed in Tables III and IV, respectively.

From Fig. 3, both MOS and SIM values of EStarGAN outperform those of AutoVC under all evaluated conditions. These results demonstrate that the proposed EStarGAN system can provide reasonable sound quality and similarity for converting both male and female speakers. In addition, from Tables III and IV, it can be seen that EStarGAN again provides better MOS and SIM values than AutoVC, revealing that EStarGAN can provide superior conversion effectiveness for human listeners in noisy environments.

V. Conclusions

We proposed a novel EStarGAN VC system that integrates an SE function to handle a noisy background during VC tasks. The LSTM-based model structure was used to construct the SE function, whereas the following StarGAN architecture was used to conduct the conversion. AutoVC, which has been proven to provide a good VC performance in a clean environment, is used as the generator in the StarGAN. The experiment results clearly indicate that the newly proposed EStarGAN VC model significantly improves the sound quality and similarity with the target speaker of the converted speech. In addition, EStarGAN has proven to be a robust VC system under various noisy environments. In the future, we plan to study EStarGAN for singing voice conversion.

References

[1] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2222–2235, 2007.

[2] R. Aihara, T. Takiguchi, and Y. Ariki, "Many-to-one voice conversion using exemplar-based sparse representation," in *Proc. WASPAA*, pp. 1–5, 2015.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.

[4] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[5] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally linear embedding for exemplar-based spectral conversion.," in *Proc. INTERSPEECH*, pp. 1652–1656, 2016.

[6] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[7] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, pp. 4869–4873, IEEE, 2015.

[8] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APSIPA*, pp. 1–6, 2016.

[9] J. Wu, D. Huang, L. Xie, and H. Li, "Denoising recurrent neural network for deep bidirectional lstm based voice conversion.," in *Proc. INTERSPEECH*, pp. 3379–3383, 2017.

[10] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*, pp. 2100–2104, IEEE, 2018.

[11] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. SLT*, pp. 266–273, IEEE, 2018.

[12] Y. Kurita, K. Kobayashi, K. Takeda, and T. Toda, "Robustness of statistical voice conversion based on direct waveform modification against background sounds.," in *Proc. INTERSPEECH*, pp. 684–688, 2019.

[13] X. Miao, M. Sun, X. Zhang, and Y. Wang, "Noise-robust voice conversion using high-quefrency boosting via sub-band cepstrum conversion and fusion," *Applied Sciences*, vol. 10, no. 1, p. 151, 2020.

[14] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. SLT*, pp. 313–317, 2012.

[15] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Noisy-to-noisy voice conversion framework with denoising model," *arXiv preprint arXiv:2109.10608*, 2021.

[16] D. Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[17] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2016.

[18] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust ASR," in *Proc. LVA/ICA*, pp. 91–99, 2015.

[19] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[20] M. Kolboek, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Proc. SLT*, pp. 305–311, IEEE, 2016.

[21] S. Shon, H. Tang, and J. Glass, "Voiceid loss: Speech enhancement for speaker verification," *arXiv preprint arXiv:1904.03601*, 2019.

[22] K. A Al-Karawi, A. H Al-Noori, F. F. Li, T. Ritchings, *et al.*, "Automatic speaker recognition system in adverse conditions—implication of noise and reverberation on system performance," *International Journal of Information and Electronics Engineering*, vol. 5, no. 6, pp. 423–427, 2015.

[23] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[24] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder.," in *Proc. Interspeech*, vol. 2013, pp. 436–440, 2013.

[25] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[26] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *Proc. ICML*, PMLR, 2019.

[27] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.

[28] M.-W. Huang, "Development of Taiwan Mandarin hearing in noise test," *Master thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health science*, 2005.

[29] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[30] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[31] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, pp. 6199–6203, 2020.

[32] C.-y. Tseng, Y.-C. Cheng, and C. Chang, "Sinica cospro and toolkit—corpora and platform of mandarin chinese fluent speech," in *Proc. Oriental COCOSDA*, pp. 23–28, 2005.