Speech Enhancement 모델 정리

Speech Enhancement

Improve the intelligibility and quality of speech contaminated by additive noise

SI(SV, SC), SR

SEGAN: Speech Enhancement Generative Adversarial Network (17.03.)

GAN Training

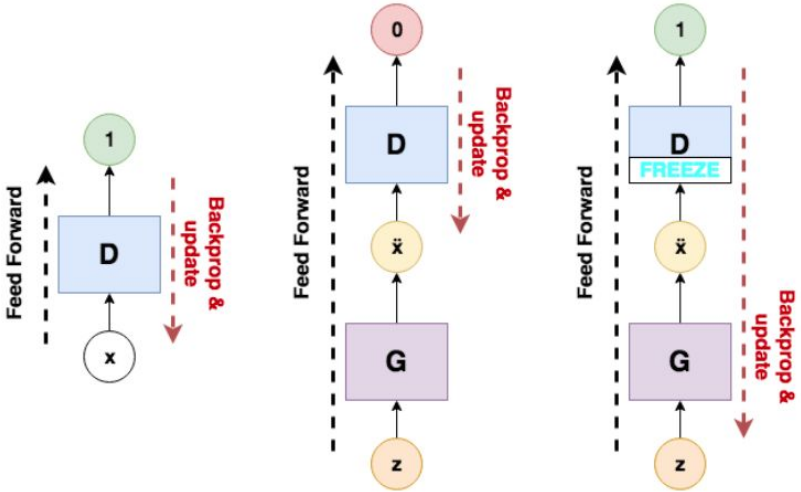

Figure 1: *GAN training process. First, D back-props a batch of real examples. Then, D back-props a batch of fake examples that come from G, and classifies them as fake. Finally, D's parameters are frozen and G back-props to make D misclassify.*

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\left[\log D(\mathbf{x})\right] + \\ + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[\log\left(1 - D\left(G\left(\mathbf{z}\right)\right)\right)\right]. \tag{1}$$

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x},\mathbf{x}_c\sim p_{\text{data}}(\mathbf{x},\mathbf{x}_c)}\left[\log D(\mathbf{x},\mathbf{x}_c)\right] + \\ + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\mathbf{x}_c\sim p_{\text{data}}(\mathbf{x}_c)}\left[\log\left(1 - D\left(G\left(\mathbf{z},\mathbf{x}_c\right),\mathbf{x}_c\right)\right)\right]. \tag{2}$$

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{x}_c\sim p_{\text{data}}(\mathbf{x},\mathbf{x}_c)}[(D(\mathbf{x},\mathbf{x}_c)-1)^2] + \\ + \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\mathbf{x}_c\sim p_{\text{data}}(\mathbf{x}_c)}[D(G(\mathbf{z},\mathbf{x}_c),\mathbf{x}_c)^2] \tag{3}$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\mathbf{x}_c\sim p_{\text{data}}(\mathbf{x}_c)}[(D(G(\mathbf{z},\mathbf{x}_c),\mathbf{x}_c)-1)^2]. \tag{4}$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}\sim p_{\text{data}}(\tilde{\mathbf{x}})}[(D(G(\mathbf{z},\tilde{\mathbf{x}}),\tilde{\mathbf{x}})-1)^2] + \\ + \lambda\left\|G(\mathbf{z},\tilde{\mathbf{x}}) - \mathbf{x}\right\|_1. \tag{5}$$

Generator

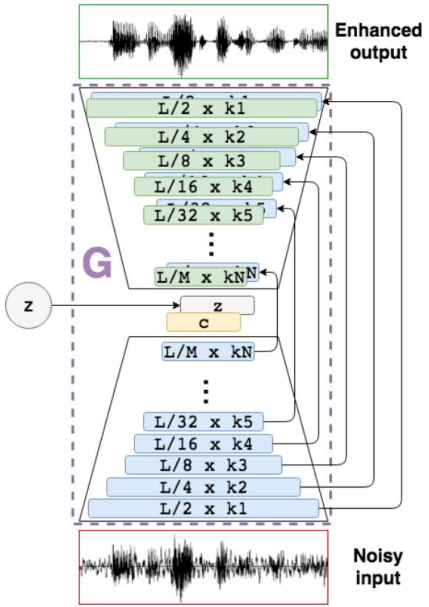

Figure 2: *Encoder-decoder architecture for speech enhancement (G network). The arrows between encoder and decoder blocks denote skip connections.*

SEGAN Training

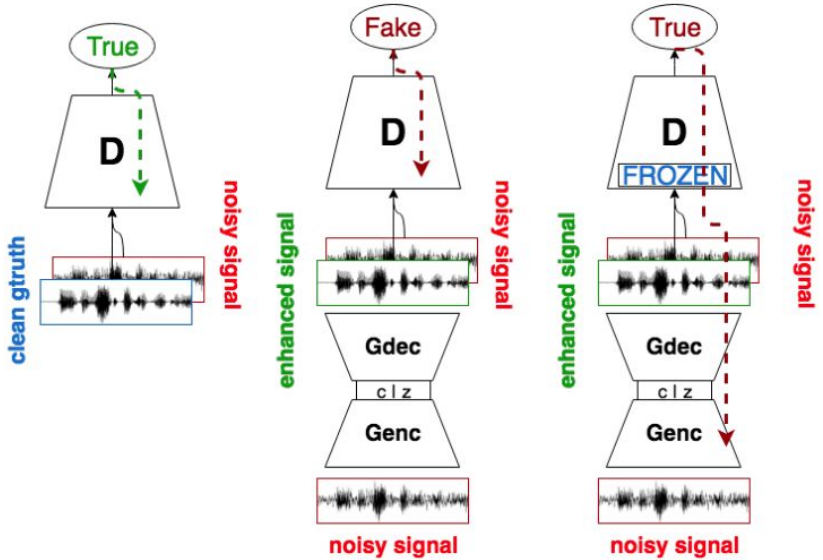

Figure 3: *Adversarial training for speech enhancement. Dashed lines represent gradient backprop.*

Generative Adversarial Network-based Postfilter for STFT Spectrograms (17.08. INTERSPEECH)

Points

TTS enhancement (NOT Speech Enhancemenet Per Se)

Language and Noise Transfer in Speech Enhancement Generative Adversarial Network (17.12.)

Reducing Over-smoothness in Speech Synthesis using Generative Adversarial Networks (18.12.)

A New GAN-based End-to-End TTS Training Algorithm (19.04. MS)

Perceptual Speech Enhancement via Generative Adversarial Networks (19.10. ICASSP)

Transformation of Low-Quality Device-Recorded Speech to High-Quality Speech Using Improved SEGAN Model (19.11.)

Improving GANs for Speech Enhancement (20.01)

SEGAN

The **single-stage** enhancement mapping
via a **single** generator G

Proposal 1: ISEGAN

The multi-stage enhancement mapping

Generators sharing parameters

Performing the same enhancement mapping iteratively

Proposal 2: DSEGAN*

The multi-stage enhancement mapping

Generators are independent

Performing the different enhancement mappings

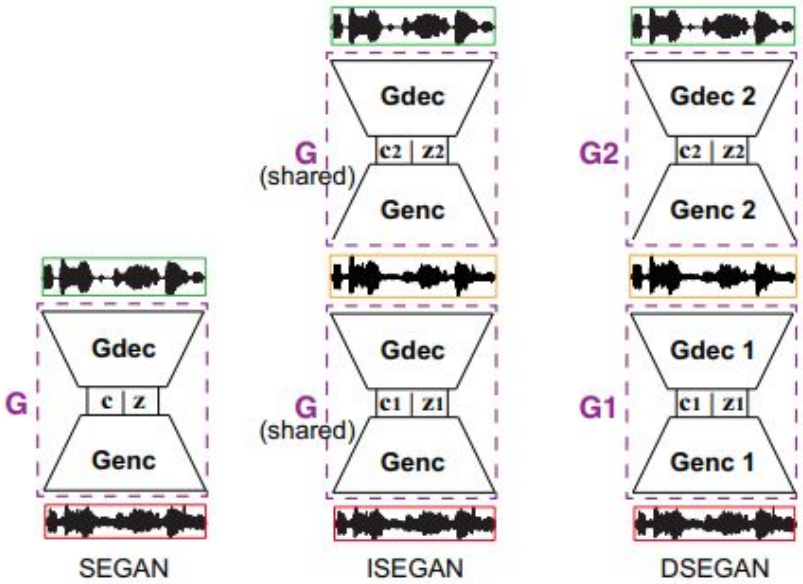

Fig. 1. Illustration of the vanilla SEGAN [14], the proposed ISEGAN with two shared generators G, and the proposed DSEGAN with two independent generators $G1$ and $G2$.

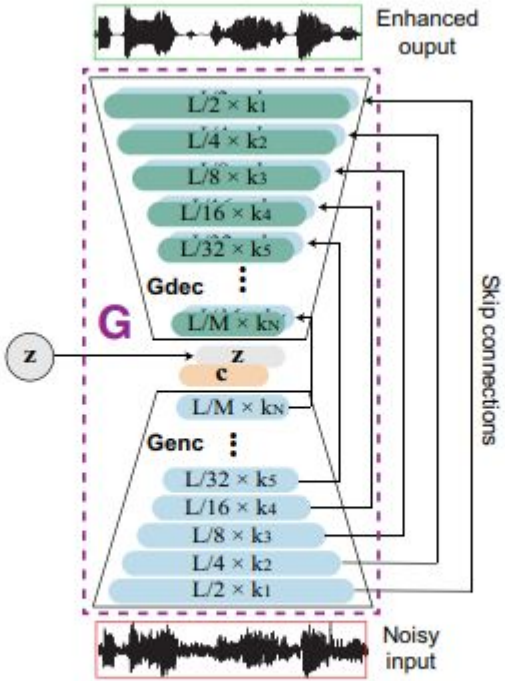

Fig. 3. The architecture of a generator in ISEGAN and DSEGAN, which is similar to the vanilla SEGAN's generator [14].

Input

Raw waveform

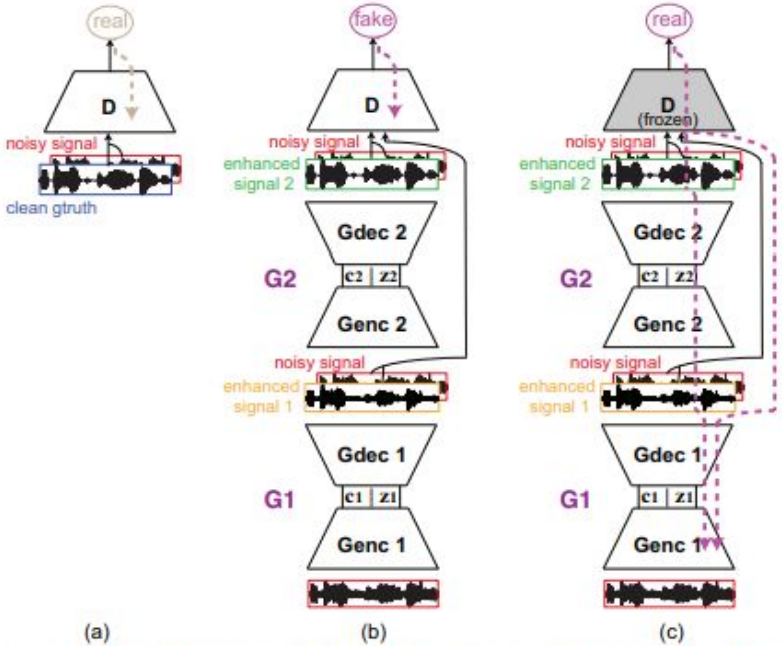

Fig. 2. Adversarial training of ISEGAN and DSEGAN. The discriminator D is learned to classify the pair $(\mathbf{x}, \tilde{\mathbf{x}})$ as real (a), and all the pairs $(\hat{\mathbf{x}}_1, \tilde{\mathbf{x}})$, $(\hat{\mathbf{x}}_2, \tilde{\mathbf{x}})$, ..., $(\hat{\mathbf{x}}_N, \tilde{\mathbf{x}})$ as fake (b). The chained generators $\mathfrak{G}$ are learned to fool D so that D classifies the pairs $(\hat{\mathbf{x}}_1, \tilde{\mathbf{x}})$, $(\hat{\mathbf{x}}_2, \tilde{\mathbf{x}})$, ..., $(\hat{\mathbf{x}}_N, \tilde{\mathbf{x}})$ as real (c). Dashed lines represent the flow of gradient backdrop.

Result

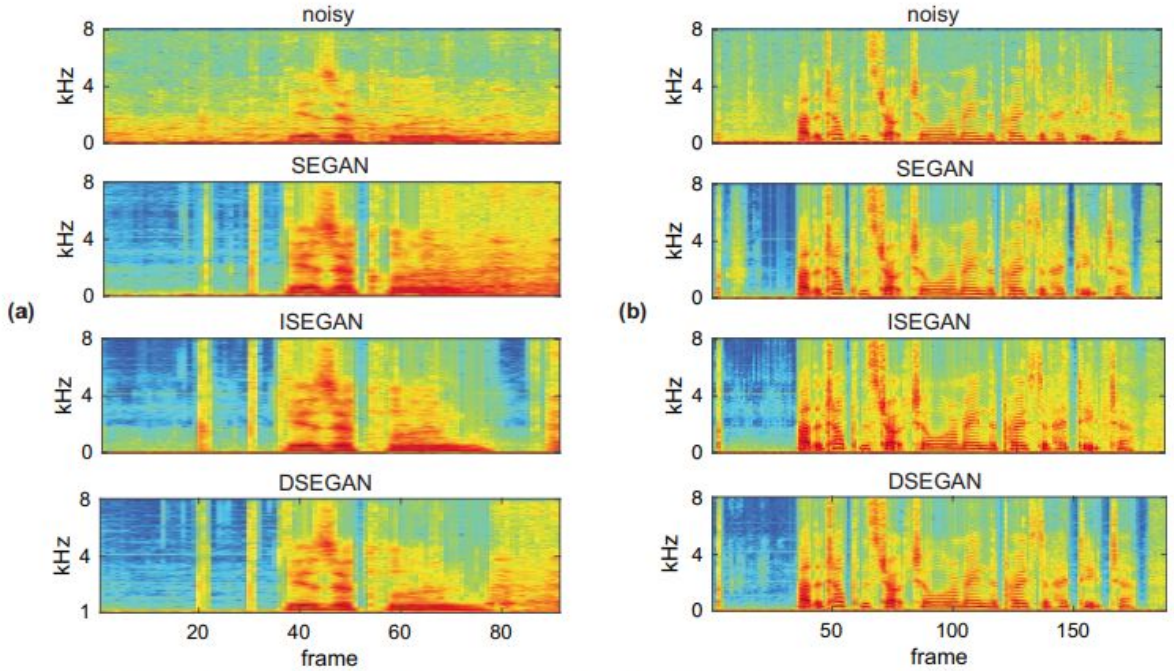

Fig. 4. Noisy signals and the enhanced signals of two test utterances produced by the vanilla SEGAN baseline, ISEGAN, and DSEGAN. (a) *p257_219.wav* and (b) *232_051.wav*.