Generative Adversarial Network

and its Applications to Signal Processing
and Natural Language Processing

Part II: Speech Signal Processing

Outline of Part II

Speech Signal Generation

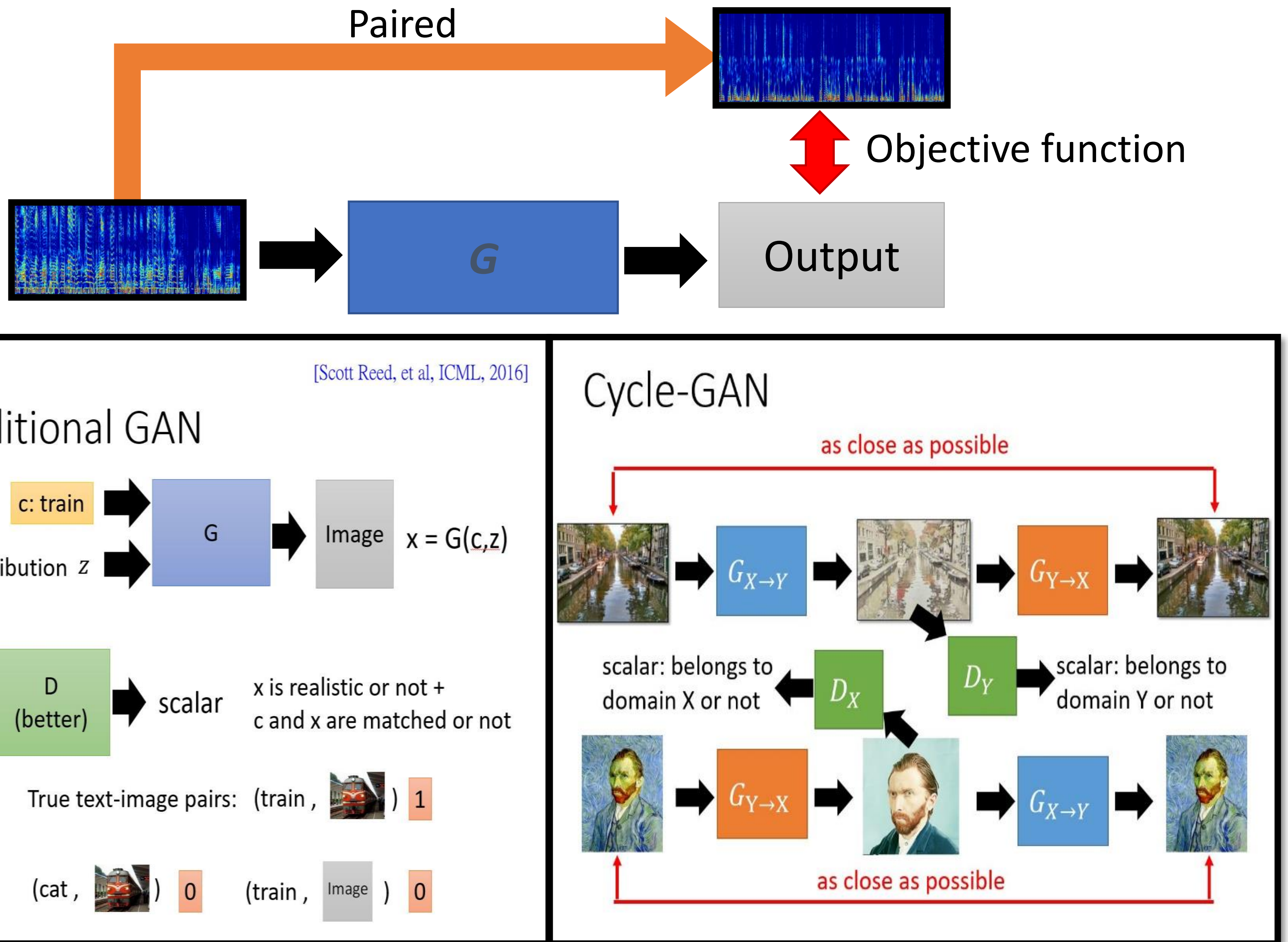
- Speech enhancement
- Postfilter, speech synthesis, voice conversion

Speech Signal Recognition

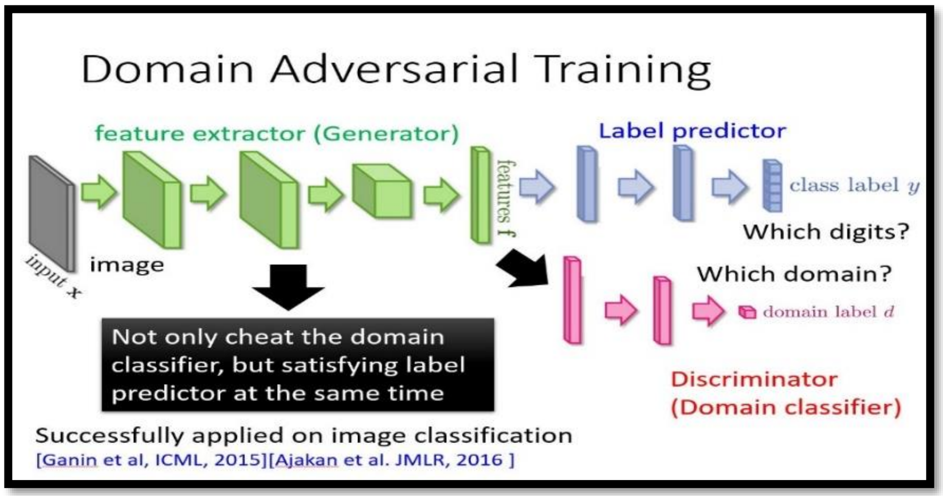
- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

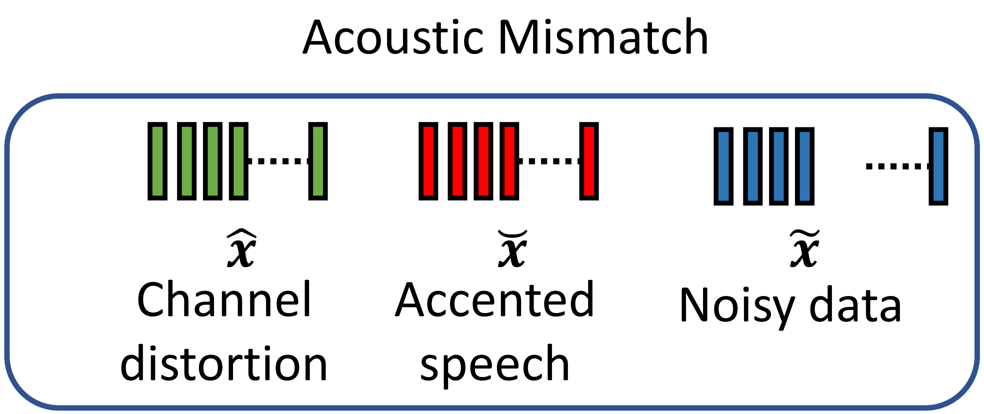
Conclusion

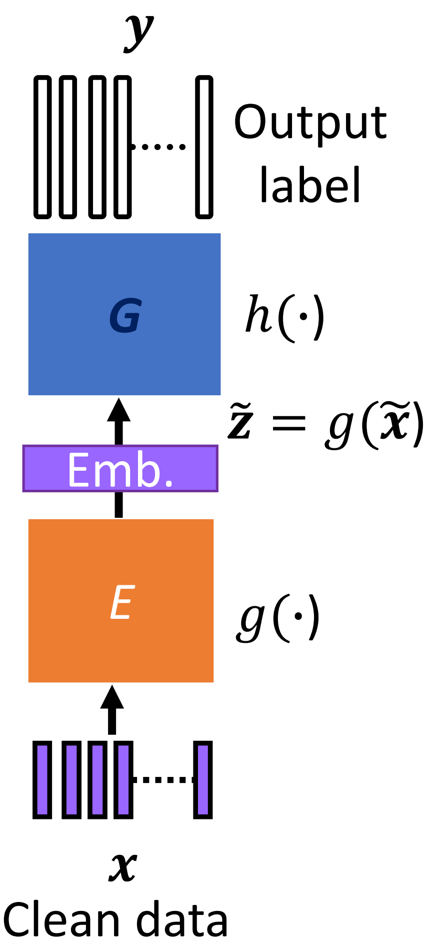
Speech Signal Generation (Regression Task)



Speech, Speaker, Emotion Recognition and Lip-reading (Classification Task)







Outline of Part II

Speech Signal Generation

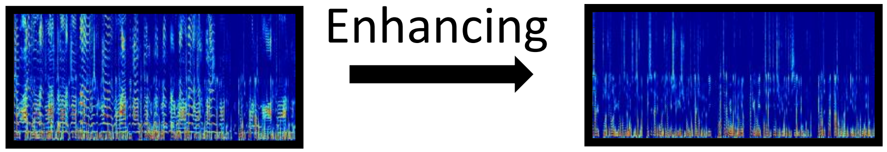
- Speech enhancement
- Postfilter, speech synthesis, voice conversion

Speech Signal Recognition

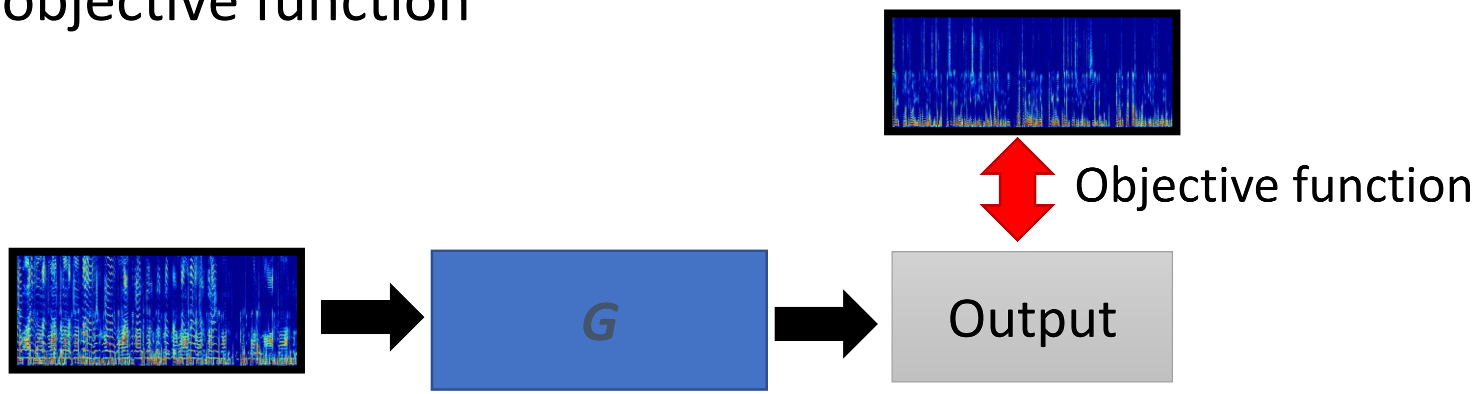
- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion

Speech Enhancement



- Typical objective function

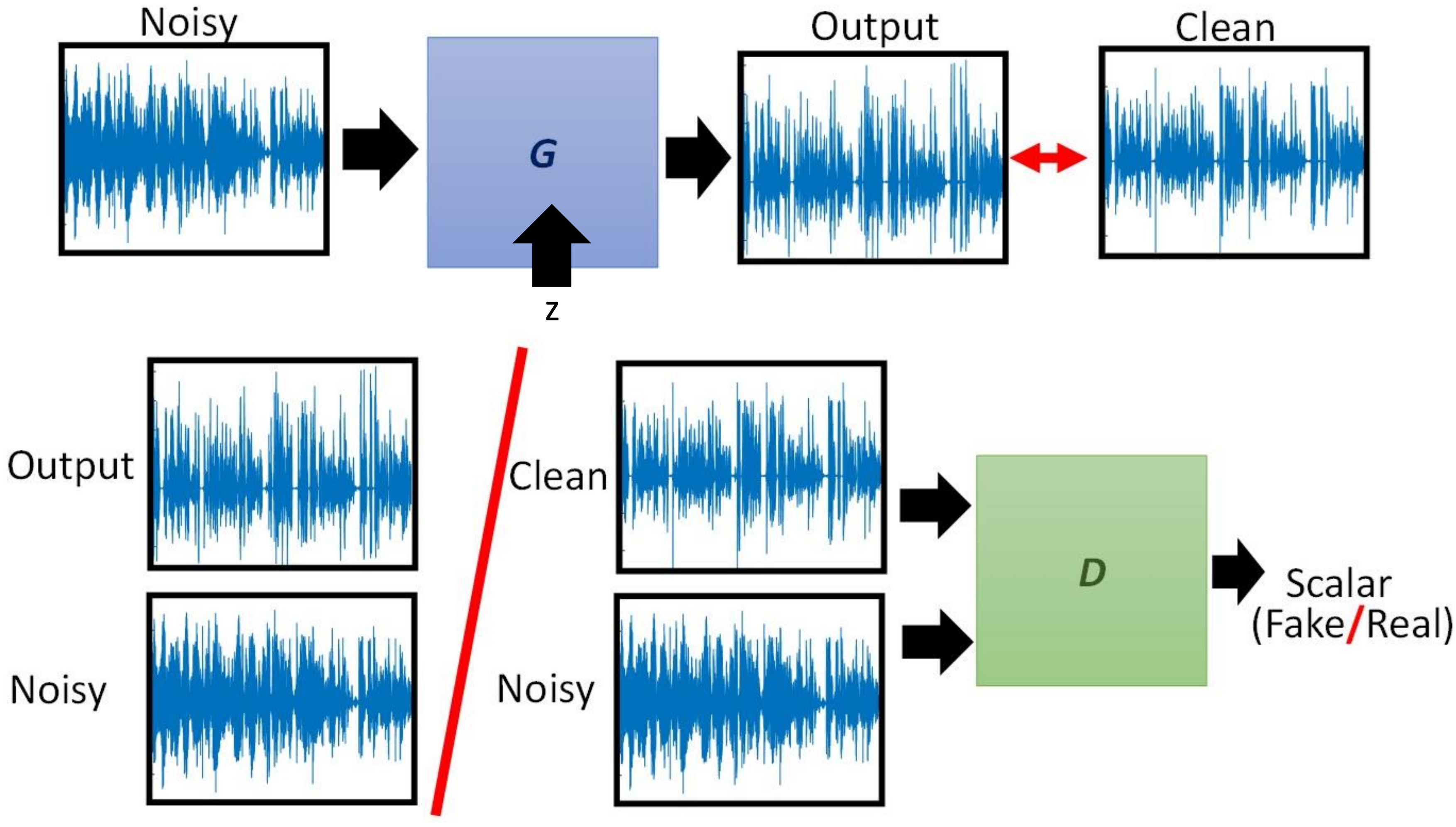


  - ➢Model structures of ***G***: DNN [Wang et al. NIPS 2012; Xu et al., SPL 2014], DDAE [Lu et al., Interspeech 2013], RNN (LSTM) [Chen et al., Interspeech 2015; Weninger et al., LVA/ICA 2015], CNN [Fu et al., Interspeech 2016].

- Typical objective function
  - ➢ Mean square error (MSE) [Xu et al., TASLP 2015], L1 [Pascual et al., Interspeech 2017], likelihood [Chai et al., MLSP 2017], STOI [Fu et al., TASLP 2018].
  - ➢ GAN is used as a new objective function to estimate the parameters in ***G***.

Speech Enhancement

- Speech enhancement GAN (SEGAN) [Pascual et al., Interspeech 2017]



Speech Enhancement (SEGAN)

- Experimental results

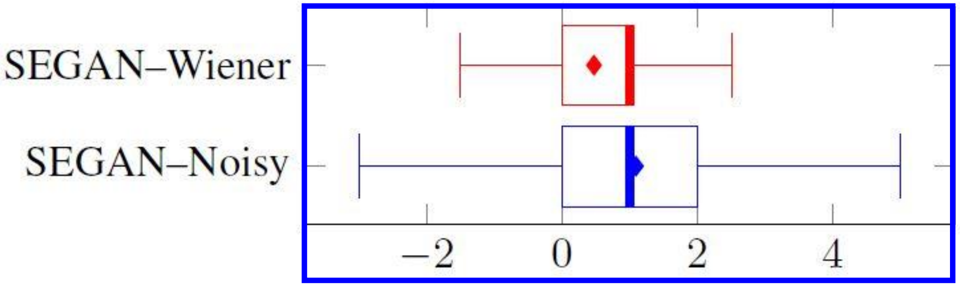
Table 1: Objective evaluation results.

Metric	Noisy	Wiener	SEGAN
PESQ	1.97	2.22	2.16
CSIG	3.35	3.23	3.48
CBAK	2.44	2.68	2.94
COVL	2.63	2.67	2.80
SSNR	1.68	5.07	7.73

Table 2: Subjective evaluation results.

Metric	Noisy	Wiener	SEGAN
MOS	2.09	2.70	3.18

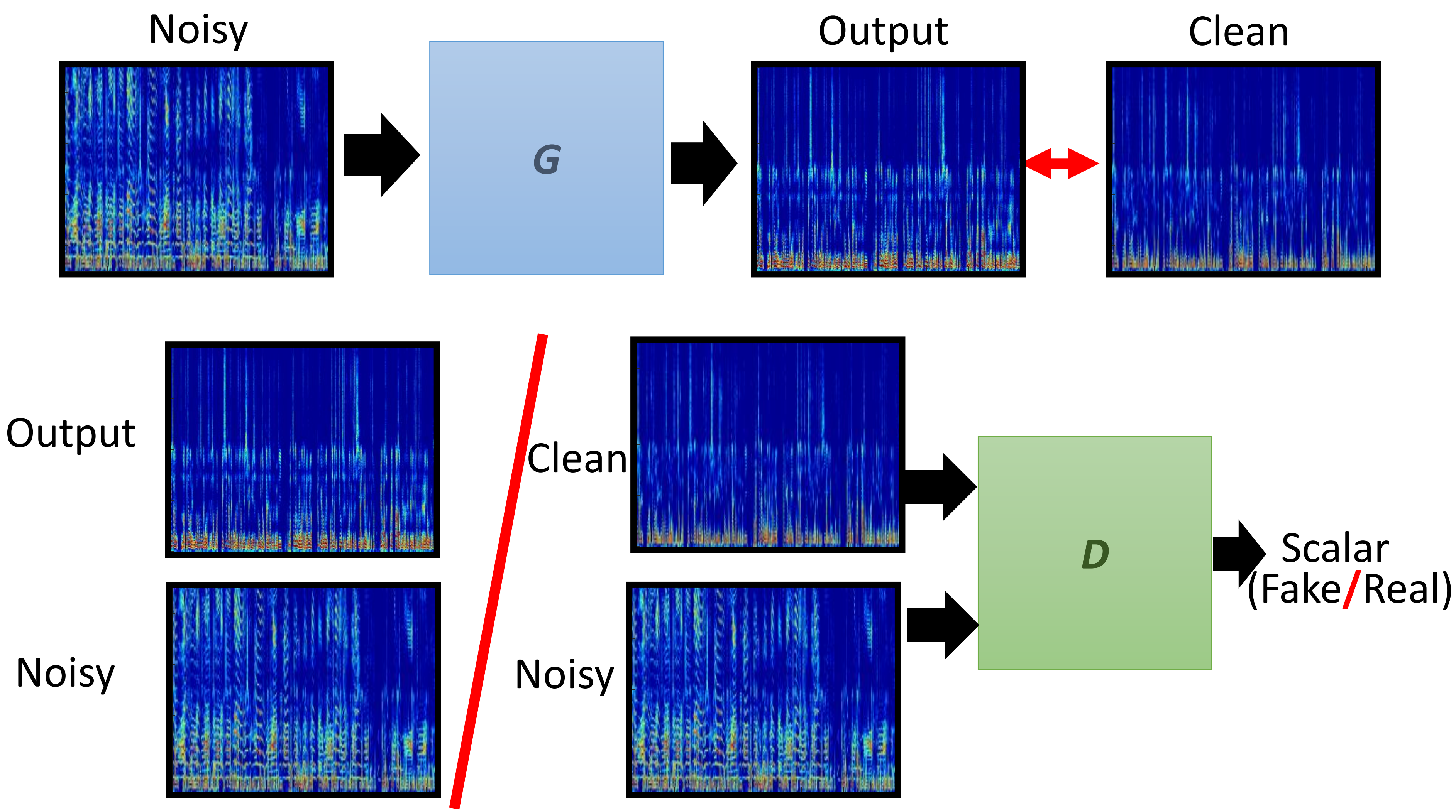
Fig. 1: Preference test results.



SEGAN yields better speech enhancement results than Noisy and Wiener.

Speech Enhancement

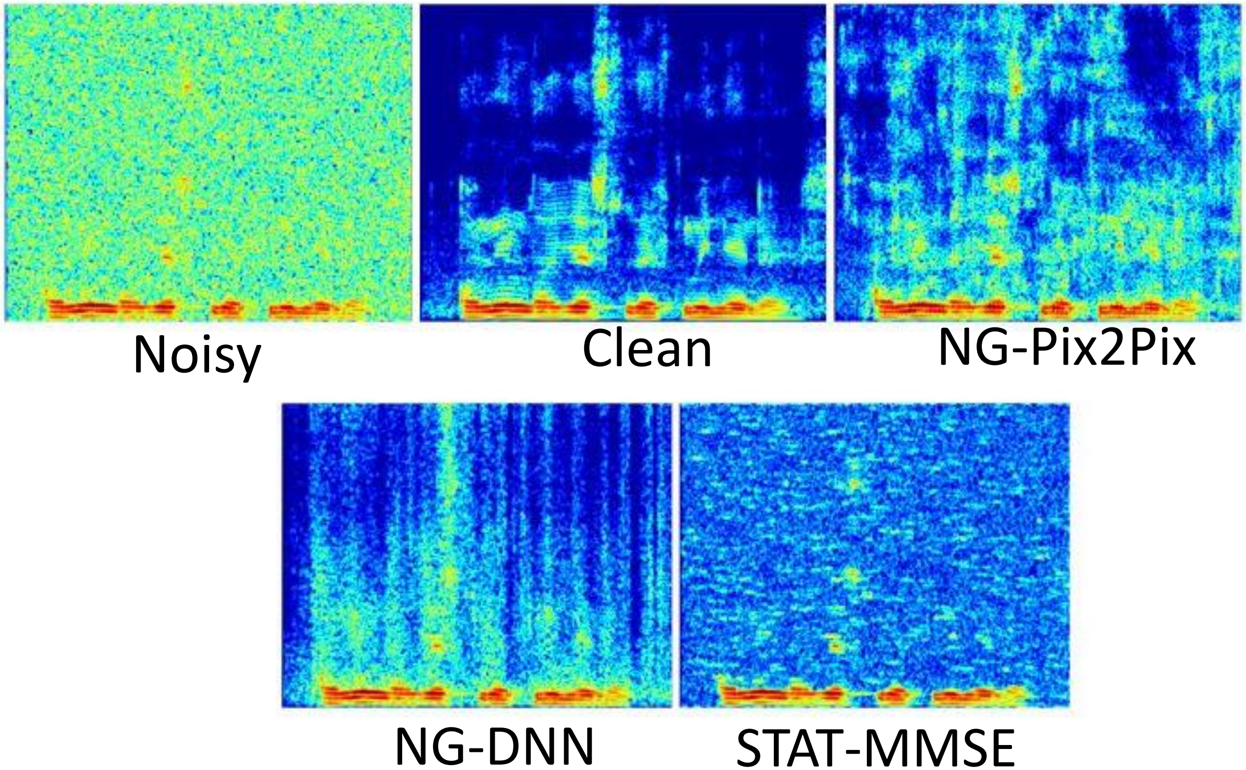
- Pix2Pix [Michelsanti et al., Interpsech 2017]



Speech Enhancement (Pix2Pix)

- Spectrogram analysis

Fig. 2: Spectrogram comparison of Pix2Pix with baseline methods.



Pix2Pix outperforms STAT-MMSE and is competitive to DNN SE.

Speech Enhancement (Pix2Pix)

- Objective evaluation and speaker verification test

Table 3: Objective evaluation results.

	PESQ						
	SNR	0	5	10	15	20	mean
Babble	(a)	1.20	1.42	1.79	2.40	**3.13**	1.99
	(b)	1.14	1.31	1.61	2.07	2.65	1.76
	(c)	**1.25**	1.51	1.87	2.31	2.78	1.95
	(d)	1.20	1.48	1.98	2.52	2.93	2.02
	(e)	1.24	**1.52**	1.88	2.31	2.78	1.95
	(f)	1.20	1.49	**2.00**	**2.53**	2.93	**2.03**

	STOI						
		0	5	10	15	20	mean
Babble	(a)	0.44	0.56	0.67	0.77	0.85	0.66
	(b)	0.43	0.56	0.66	0.74	0.81	0.64
	(c)	**0.50**	**0.63**	**0.72**	**0.79**	**0.86**	**0.70**
	(d)	0.46	0.59	0.71	0.78	0.83	0.67
	(e)	0.49	0.62	**0.72**	**0.79**	0.85	**0.70**
	(f)	0.46	0.60	0.71	0.77	0.82	0.67

Table 4: Speaker verification results.

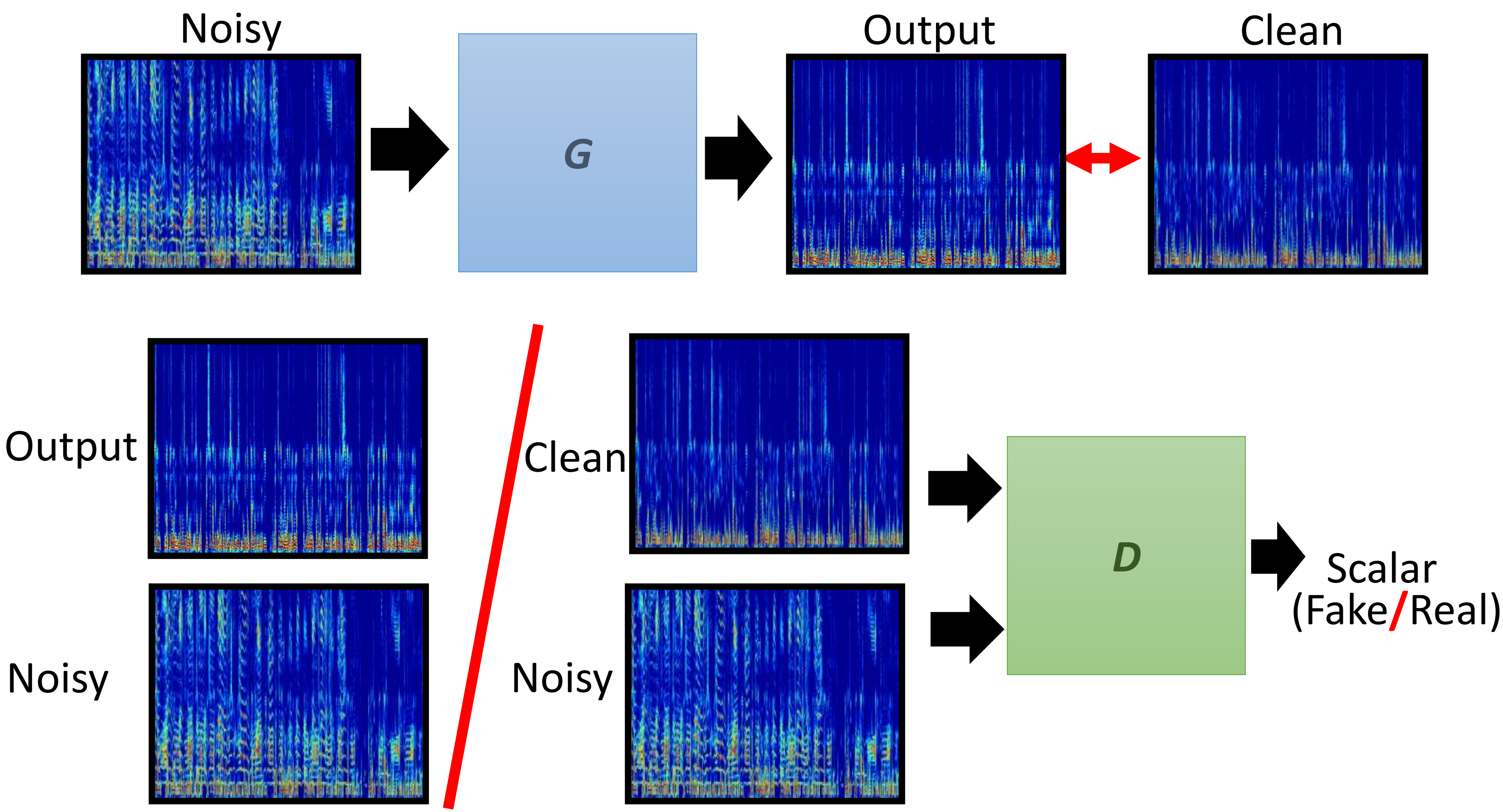
	SNR	0	5	10	15	20	clean	mean
Airplane	(a)	21.09	15.99	13.61	11.66	9.18	6.99	13.08
	(b)	17.69	12.58	8.17	6.53	6.27	5.80	9.51
	(c)	16.99	10.55	7.48	6.99	6.15	6.12	9.05
	(d)	17.19	8.84	**5.44**	5.05	**4.63**	**3.74**	7.48
	(e)	15.99	8.99	6.12	6.12	5.58	5.67	8.08
	(f)	**15.31**	**7.89**	5.58	**4.77**	4.76	5.44	**7.29**

(a)	No enhancement
(b)	STSA-MMSE
(c)	NS-DNN
(d)	**NS-Pix2Pix**
(e)	NG-DNN
(f)	**NG-Pix2Pix**

1. From the objective evaluations, Pix2Pix outperforms Noisy and MMSE and is competitive to DNN SE.
2. From the speaker verification results, Pix2Pix outperforms the baseline models when the clean training data is used.

Speech Enhancement

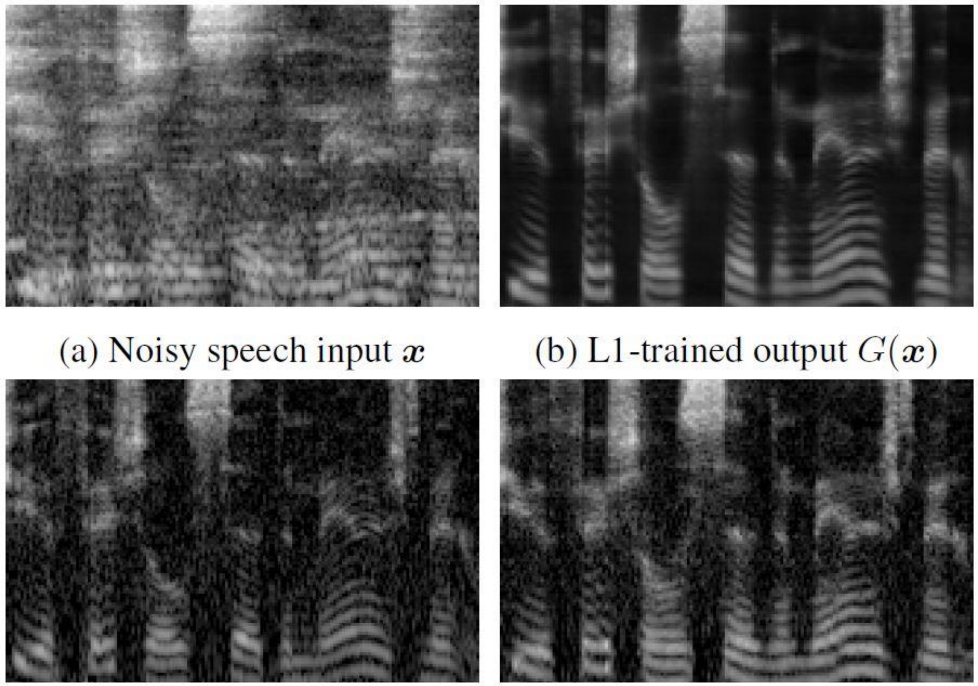
- Frequency-domain SEGAN (FSEGAN) [Donahue et al., ICASSP 2018]



Speech Enhancement (FSEGAN)

- Spectrogram analysis

Fig. 2: Spectrogram comparison of FSEGAN with L1-trained method.



(a) Noisy speech input $\boldsymbol{x}$ (b) L1-trained output $G(\boldsymbol{x})$

(c) Clean speech target $\boldsymbol{y}$ (d) FSEGAN output $G(\boldsymbol{x})$

FSEGAN reduces both additive noise and reverberant smearing.

Speech Enhancement (FSEGAN)

- ASR results

Table 5: WER (%) of SEGAN and FSEGAN.

Test Set	Enhancer	ASR-Clean WER	ASR-MTR WER
Clean	None	11.9	14.3
MTR	None	72.2	20.3
	SEGAN	80.7	52.8
	FSEGAN	33.3	25.4

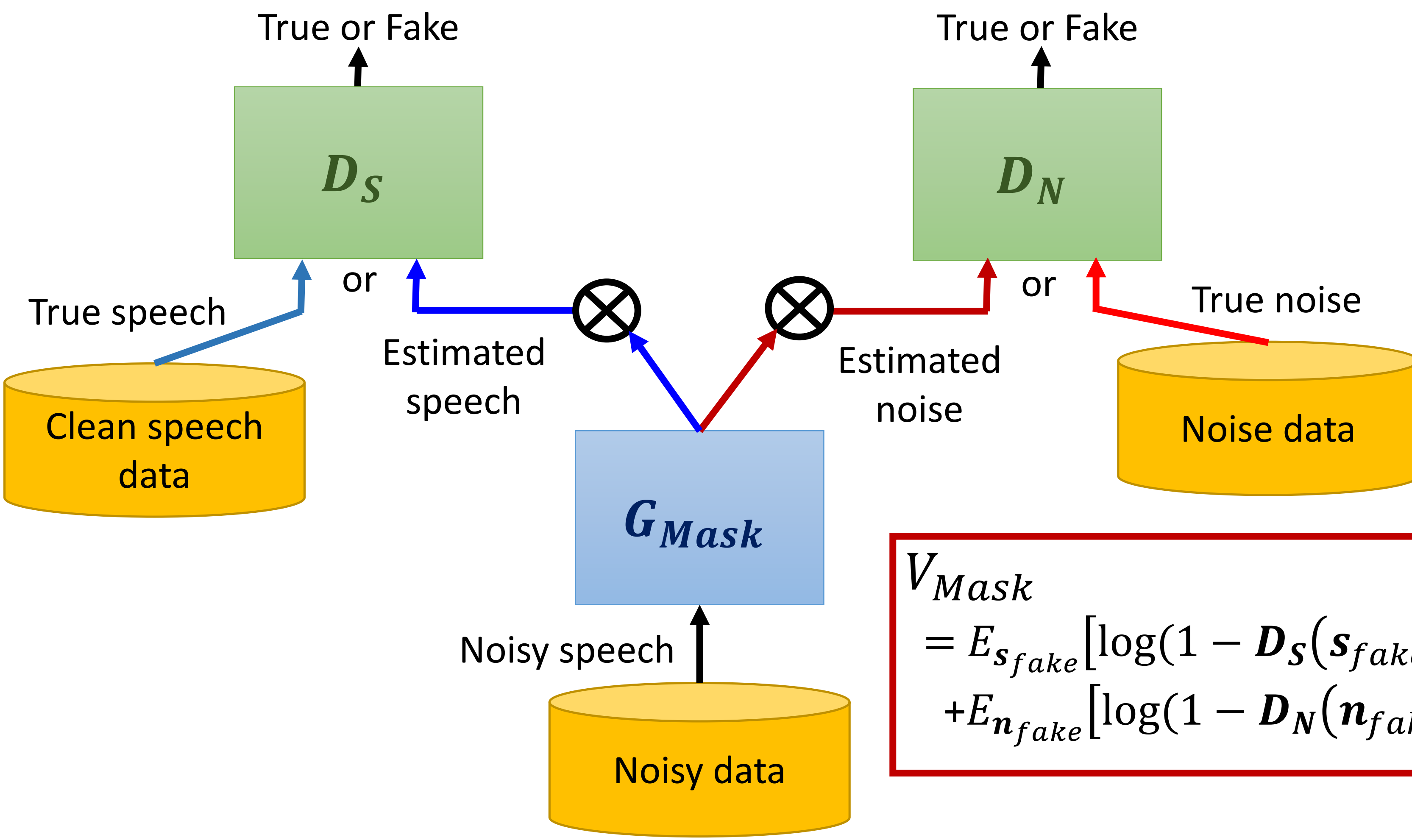
Table 6: WER (%) of FSEGAN with retrain.

Model	WER (%)
MTR Baseline *	20.3
+ Stereo	19.0
MTR + FSEGAN Enhancer *	25.4
+ Retraining	21.0
+ Hybrid Retraining	17.6
MTR + L1-trained Enhancer *	21.4
+ Retraining	18.0
+ Hybrid Retraining	17.1

1. From Table 5, (1) FSEGAN improves recognition results for ASR-Clean.
   (2) FSEGAN outperforms SEGAN as front-ends.
2. From Table 6, (1) Hybrid Retraining with FSEGAN outperforms Baseline;
   (2) FSEGAN retraining slightly underperforms L1–based retraining.

Speech Enhancement

- Adversarial training based mask estimation (ATME)
  [Higuchi et al., ASRU 2017]

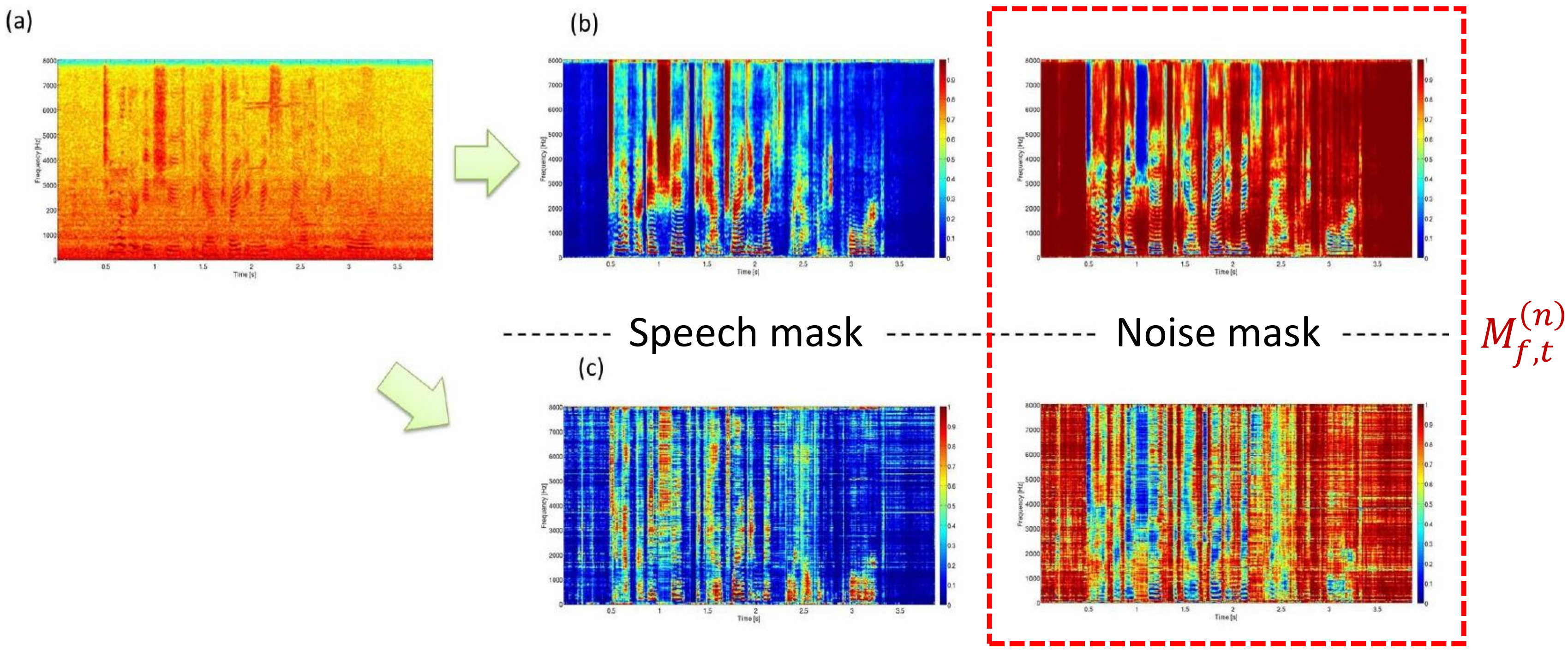


$$
\begin{aligned}
&V_{Mask} \\
&\quad = E_{\boldsymbol{s}_{fake}}\left[\log\left(1 - \boldsymbol{D_S}\left(\boldsymbol{s}_{fake}, \theta\right)\right)\right] \\
&\quad\quad + E_{\boldsymbol{n}_{fake}}\left[\log\left(1 - \boldsymbol{D_N}\left(\boldsymbol{n}_{fake}, \theta\right)\right)\right]
\end{aligned}
$$

Speech Enhancement (ATME)

- Spectrogram analysis

Fig. 3: Spectrogram comparison of (a) noisy; (b) MMSE with supervision; (c) ATMB without supervision.



The proposed adversarial training mask estimation can capture speech/noise signals without supervised data.

Speech Enhancement (ATME)

- Mask-based beamformer for robust ASR

➢ The estimated mask parameters are used to compute spatial covariance matrix for MVDR beamformer.

➢ $\hat{s}_{f,t} = \boxed{\mathbf{w}_f^{\mathrm{H}}} \mathbf{y}_{f,t}$, where $\hat{s}_{f,t}$ is the enhanced signal, and $\mathbf{y}_{f,t}$ denotes the observation of M microphones, f and t are frequency and time indices; $\mathbf{w}_f$ denotes the beamformer coefficient.

➢ The MVDR solves $\mathbf{w}_f$ by: $\mathbf{w}_f = \frac{(R_f^{(s+n)})^{-1} \boxed{\mathbf{h}_f}}{\boxed{\mathbf{h}_f^{\mathrm{H}}} (R_f^{(s+n)})^{-1} \boxed{\mathbf{h}_f}}$

➢ To estimate $\mathbf{h}_f$, the spatial covariance matrix of the target signal, $\boxed{R_f^{(s)}}$, is computed by : $\boxed{R_f^{(s)} = R_f^{(s+n)} - R_f^{(n)}}$, where $\boxed{R_f^{(n)} = \frac{M_{f,t}^{(n)} \mathbf{y}_{f,t} \mathbf{y}_{f,t}^{\mathrm{H}}}{\sum_{f,t} M_{f,t}^{(n)}}, \boxed{M_{f,t}^{(n)}}}$ was computed by AT.

Speech Enhancement (ATME)

- ASR results

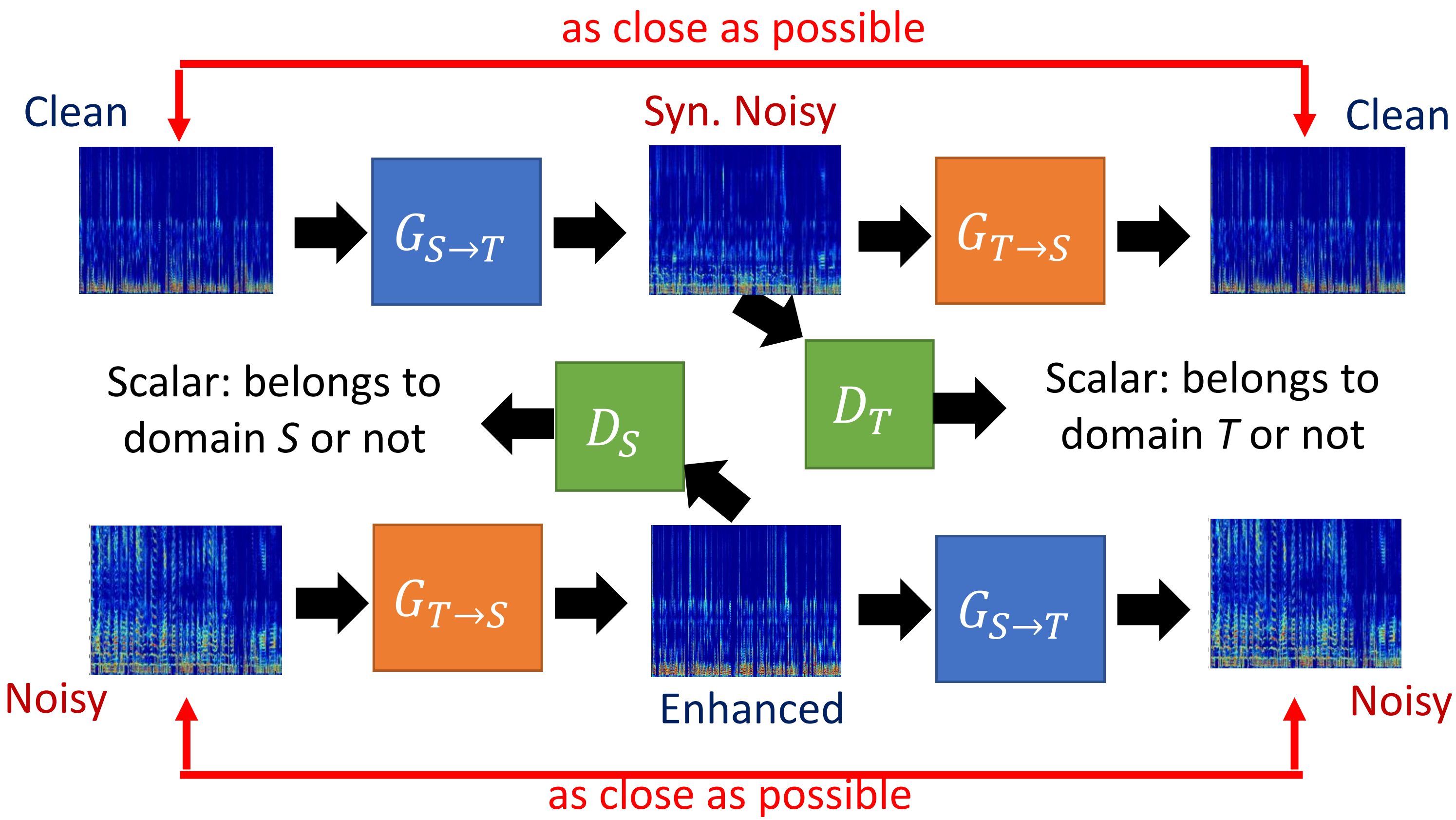
Table 7: WERs (%) for the development and evaluation sets.

systems	dev					eval				
	avg	bus	caf	ped	str	avg	bus	caf	ped	str
Unprocessed	9.01	14.00	7.94	6.03	8.05	15.60	22.55	16.21	12.89	10.74
Adversarial Training	5.00	7.60	4.09	4.03	4.29	7.58	10.24	7.51	6.20	6.39
MMSE	4.83	7.20	4.04	3.98	4.10	7.04	9.25	6.67	6.02	6.24

1. ATME provides significant improvements over Unprocessed.
2. Unsupervised ATME slightly underperforms supervised MMSE.

Speech Enhancement (AFT)

- Cycle-GAN-based acoustic feature transformation (AFT) [Mimura et al., ASRU 2017]



$$V_{Full} = V_{GAN}(G_{X\to Y}, D_Y) + V_{GAN}(G_{X\to Y}, D_Y) + \lambda\, V_{Cyc}(G_{X\to Y}, G_{Y\to X})$$

Speech Enhancement (AFT)

- ASR results on noise robustness and style adaptation

Table 8: Noise robust ASR.

acoustic model	feature	cycle loss	λ and μ	WER	ID
no adapt.	no adapt.	-	-	41.08	(1)
no adapt.	adapt. with $G_{T\to S}$	no	1, 1	55.45	(2)
		yes	1, 1	37.34	(3)
		yes	trained	36.56	(4)
adapt. with $G_{S\to T}$	no adapt.	yes	1, 1	35.98	(5)
		yes	trained	34.31	(6)

S: Clean; T: Noisy

Table 9: Speaker style adaptation.

source	target	feature	WER
JNAS	CSJ-SPS	no adapt.	26.47
		adapt. with $G_{T\to S}$	25.93
CSJ-APS	CSJ-SPS	no adapt.	17.15
		adapt. with $G_{T\to S}$	16.60

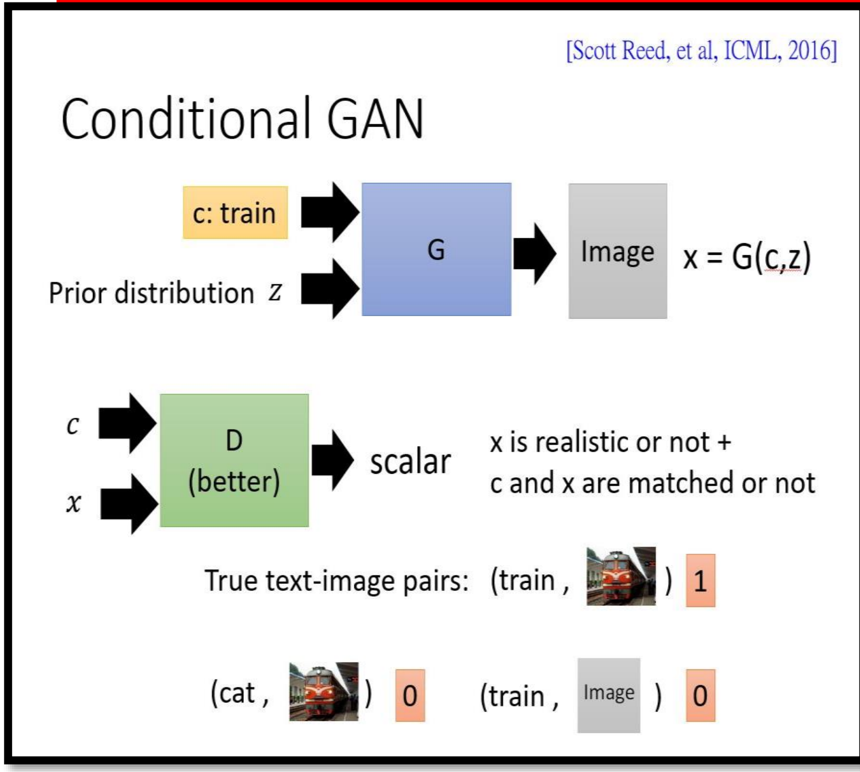
JNAS: Read; CSJ-SPS: Spontaneous (relax);
CSJ-APS: Spontaneous (formal);

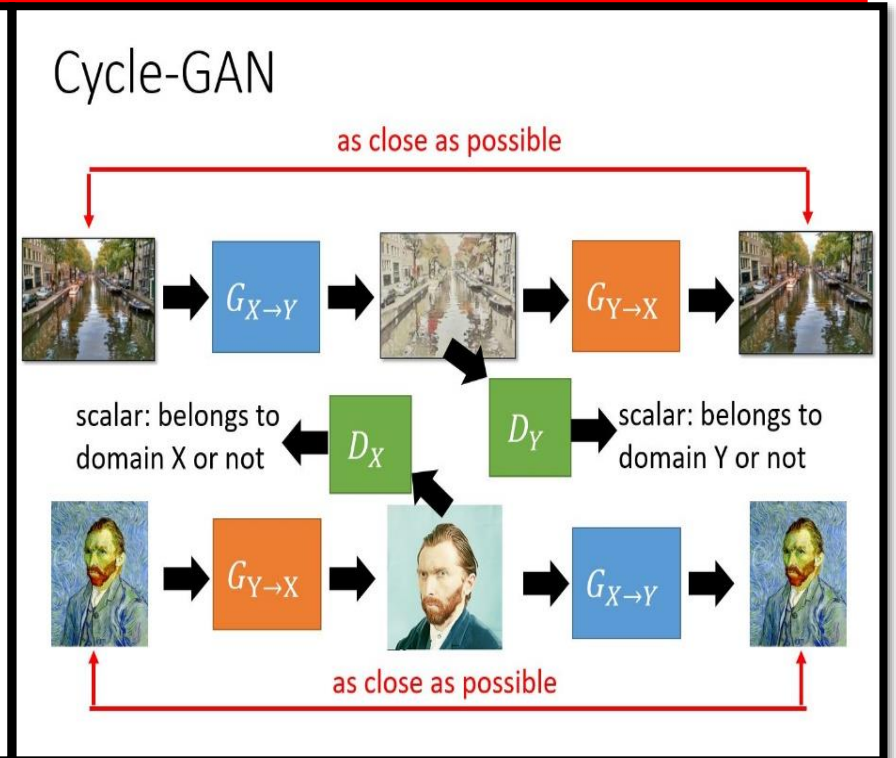
1. $G_{T\to S}$ can transform acoustic features and effectively improve ASR results for both noisy and accented speech.
2. $G_{S\to T}$ can be used for model adaptation and effectively improve ASR results for noisy speech.

Outline of Part II

Speech Signal Generation

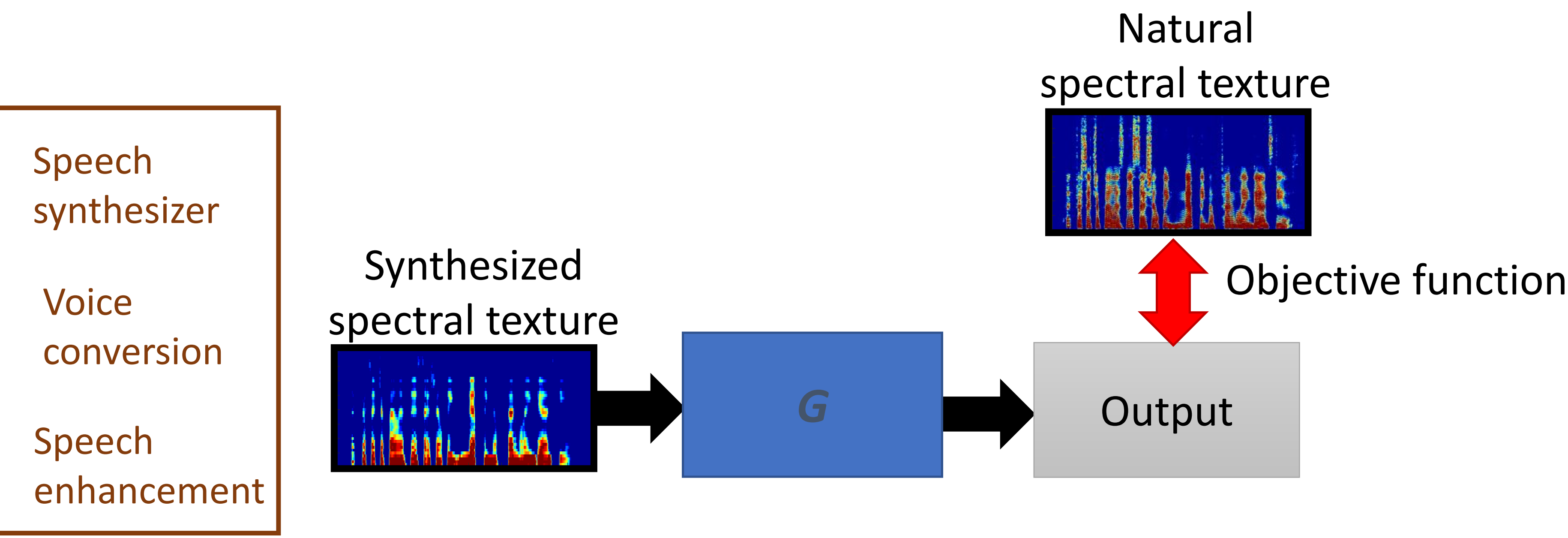
- Speech enhancement
- Postfilter, speech synthesis, voice conversion





Postfilter

- Postfilter for synthesized or transformed speech

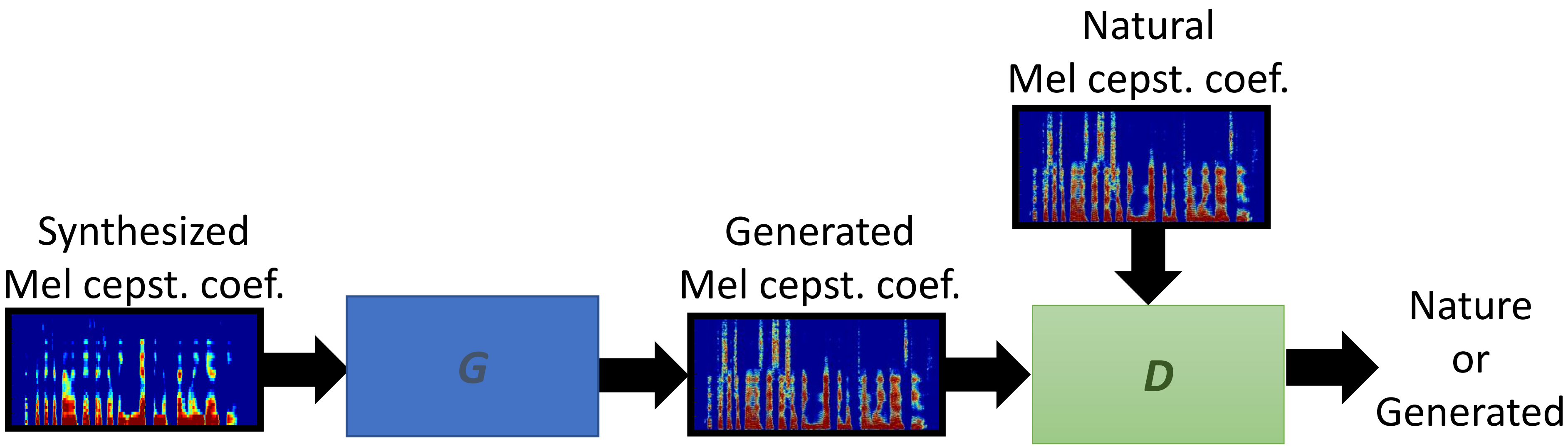


➢ Conventional postfilter approaches for ***G*** estimation include global variance (GV) [Toda et al., IEICE 2007], variance scaling (VS) [Sil’en et al., Interpseech 2012], modulation spectrum (MS) [Takamichi et al., ICASSP 2014],DNN with MSE criterion [Chen et al., Interspeech 2014; Chen et al., TASLP 2015].

➢ GAN is used a new objective function to estimate the parameters in ***G***.

Postfilter

- GAN postfilter [Kaneko et al., ICASSP 2017]

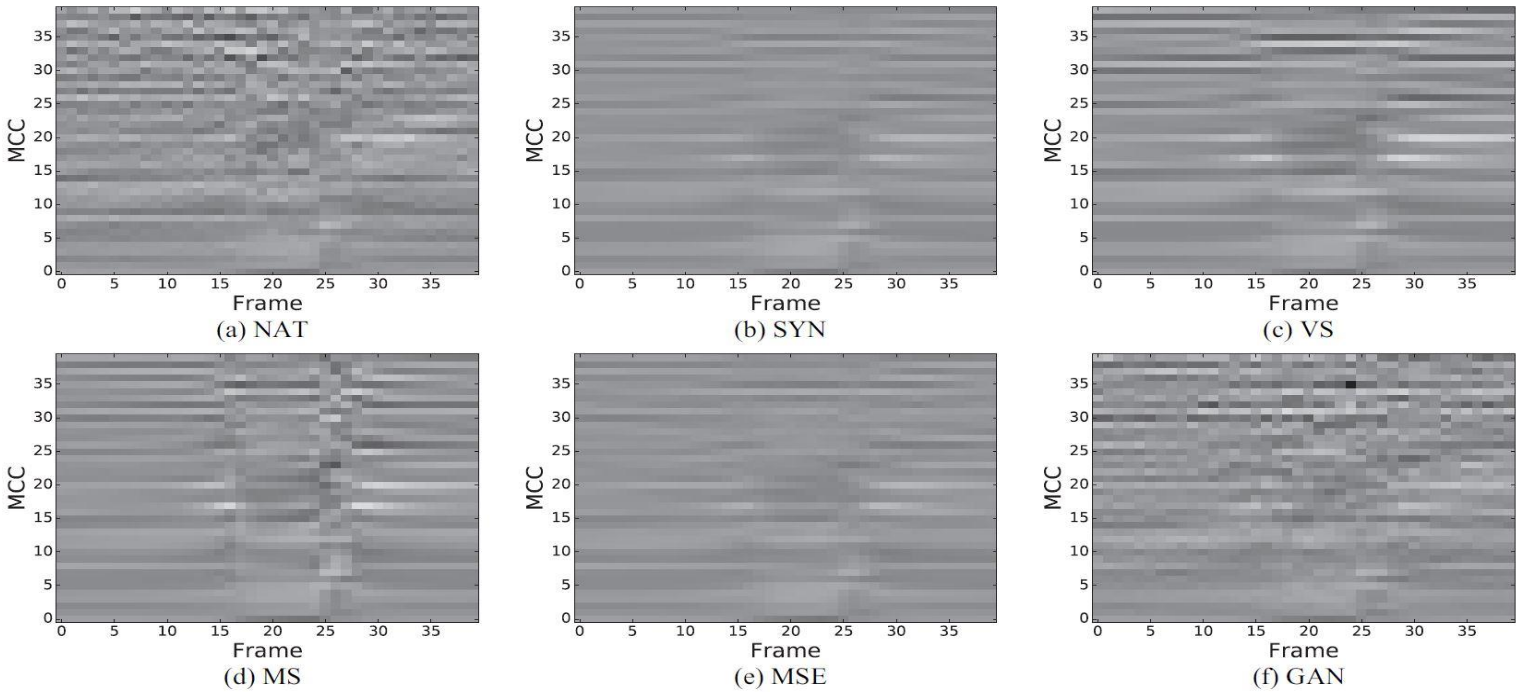


- ➢ Traditional MMSE criterion results in statistical averaging.
- ➢ GAN is used as a new objective function to estimate the parameters in ***G***.
- ➢ The proposed work intends to further improve the naturalness of synthesized speech or parameters from a synthesizer.

Postfilter (GAN-based Postfilter)

- Spectrogram analysis

Fig. 4: Spectrograms of: (a) NAT (nature); (b) SYN (synthesized); (c) VS (variance scaling); (d) MS (modulation spectrum); (e) MSE; (f) GAN postfilters.



GAN postfilter reconstructs spectral texture similar to the natural one.

Postfilter (GAN-based Postfilter)

- Objective evaluations

Fig. 5: Mel-cepstral trajectories (GANv: GAN was applied in voiced part).

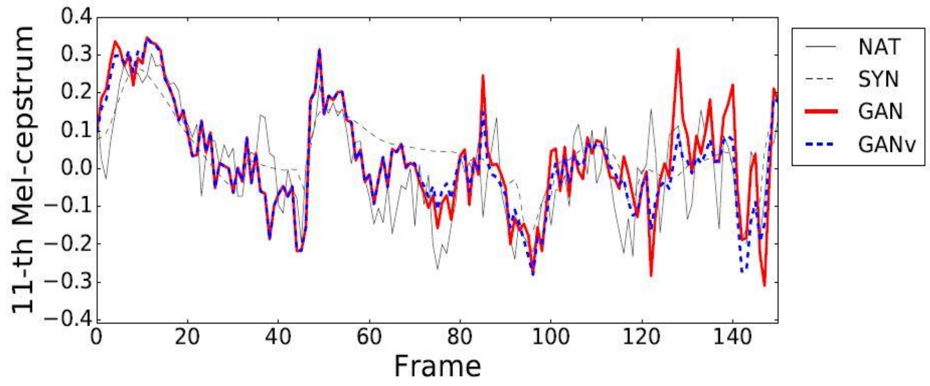
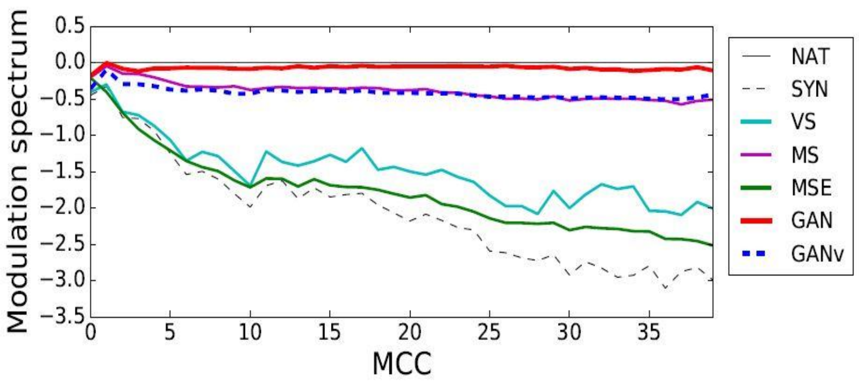


Fig. 6: Averaging difference in modulation spectrum per Mel-cepstral coefficient.



GAN postfilter reconstructs spectral texture similar to the natural one.

Postfilter (GAN-based Postfilter)

- Subjective evaluations

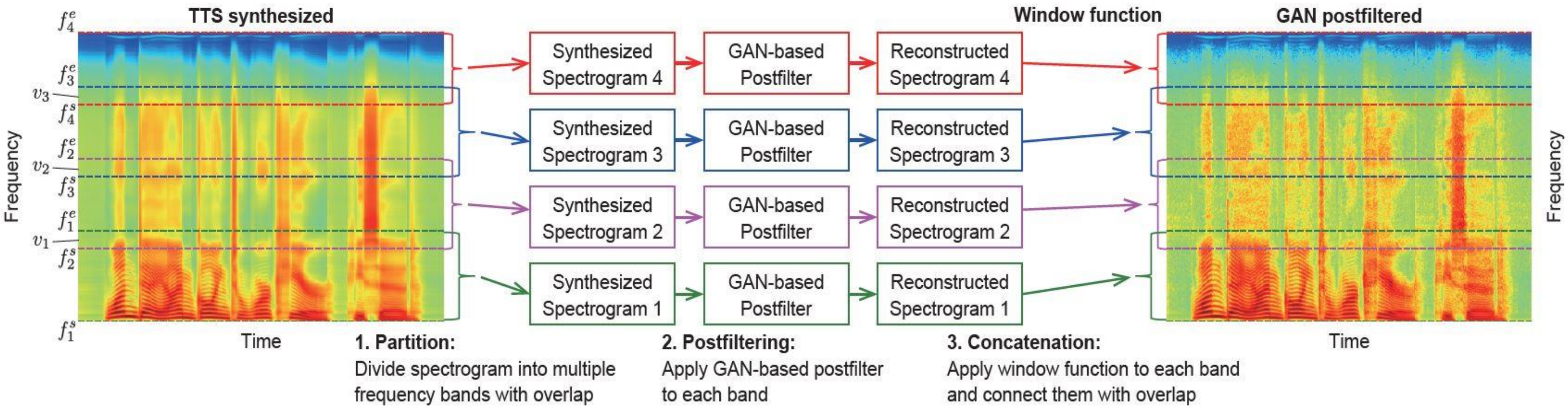
Table 10: Preference score (%). Bold font indicates the numbers over 30%.

	Former	Latter	Neutral
GAN vs. SYN	**56.5** $\pm$ 4.9	22.0 $\pm$ 4.1	21.5 $\pm$ 4.0
GAN vs. GANv	11.3 $\pm$ 3.1	**37.3** $\pm$ 4.8	**51.5** $\pm$ 4.9
GAN vs. NAT	16.8 $\pm$ 3.7	**53.5** $\pm$ 4.9	29.8 $\pm$ 4.5
GANv vs. NAT	**30.3** $\pm$ 4.5	**34.5** $\pm$ 4.7	**35.3** $\pm$ 4.7

1. GAN postfilter significantly improves the synthesized speech.
2. GAN postfilter is effective particularly in voiced segments.
3. GANv outperforms GAN and is comparable to NAT.

Postfilter (GAN-postfilter-SFTF)

- GAN post-filter for STFT spectrograms [Kaneko et al., Interspeech 2017]

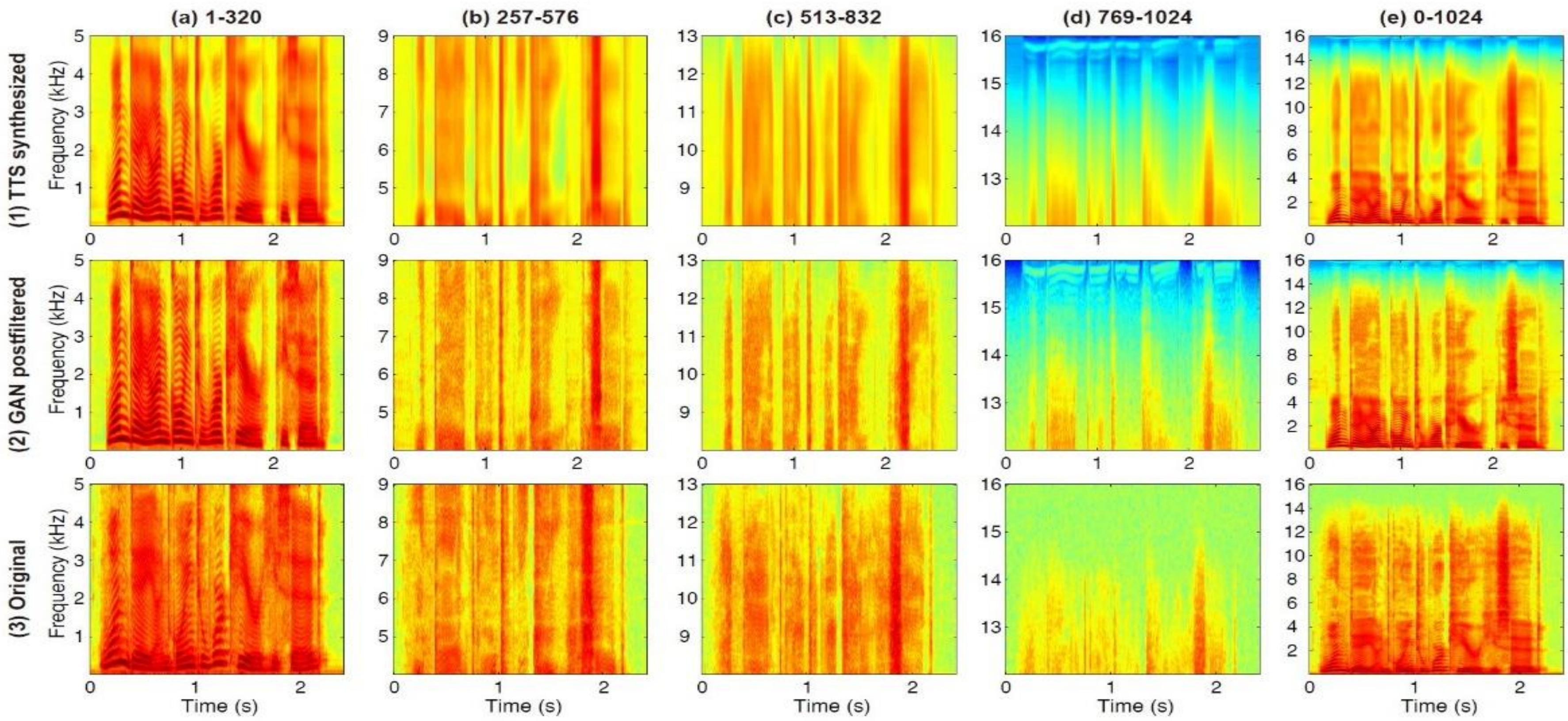


- GAN postfilter was applied on high-dimensional STFT spectrograms.
- The spectrogram was partitioned into *N* bands (each band overlaps its neighboring bands).
- The GAN-based postfilter was trained for each band.
- The reconstructed spectrogram from each band was smoothly connected.

Postfilter (GAN-postfilter-SFTF)

- Spectrogram analysis

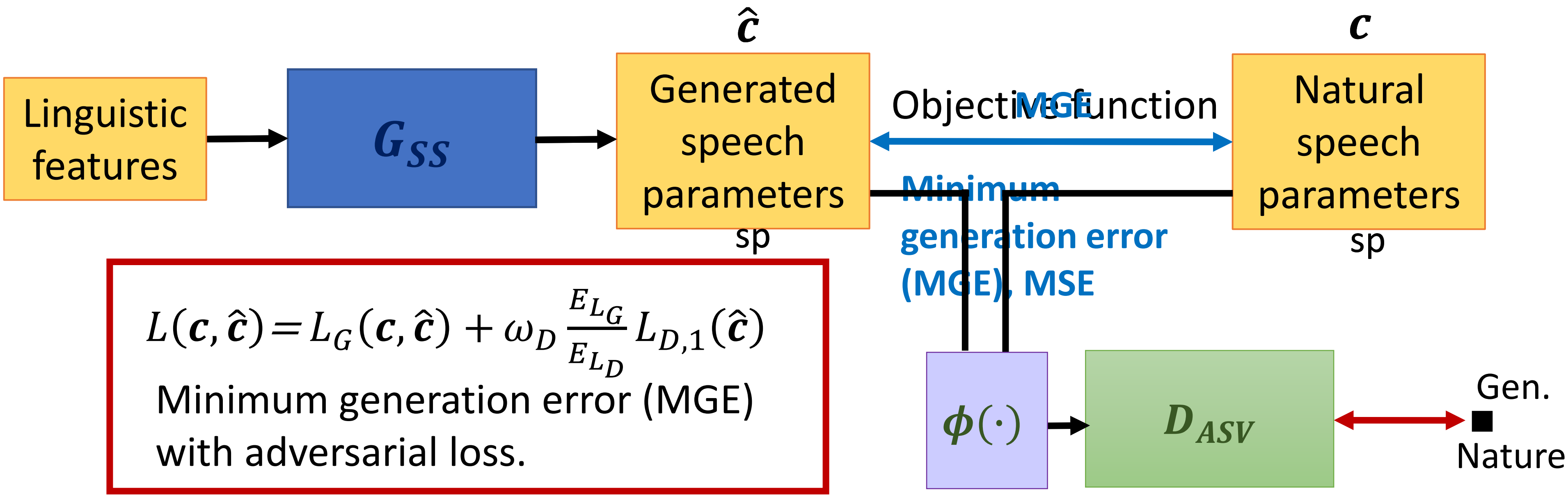
Fig. 7: Spectrograms of: (1) SYN, (2) GAN, (3) Original (NAT)



GAN postfilter reconstructs spectral texture similar to the natural one.

Speech Synthesis

- Speech synthesis with anti-spoofing verification (ASV)
  Input: linguistic features, Output: speech parameters
  [Saito et al., ICASSP 2017]



$$L(\boldsymbol{c},\hat{\boldsymbol{c}}) = L_G(\boldsymbol{c},\hat{\boldsymbol{c}}) + \omega_D \frac{E_{L_G}}{E_{L_D}} L_{D,1}(\hat{\boldsymbol{c}})$$

Minimum generation error (MGE) with adversarial loss.

$$L_D(\boldsymbol{c},\hat{\boldsymbol{c}}) = L_{D,1}(\boldsymbol{c}) + L_{D,0}(\hat{\boldsymbol{c}})$$

$$L_{D,1}(\boldsymbol{c}) = -\frac{1}{T}\sum_{t=1}^{T} \log(D(\boldsymbol{c}_t))\text{...NAT}$$

$$L_{D,0}(\hat{\boldsymbol{c}}) = -\frac{1}{T}\sum_{t=1}^{T} \log(1 - D(\hat{\boldsymbol{c}}_t))\text{...SYN}$$

Speech Synthesis (ASV)

- Objective and subjective evaluations

Fig. 8: Averaged GVs of MCCs.

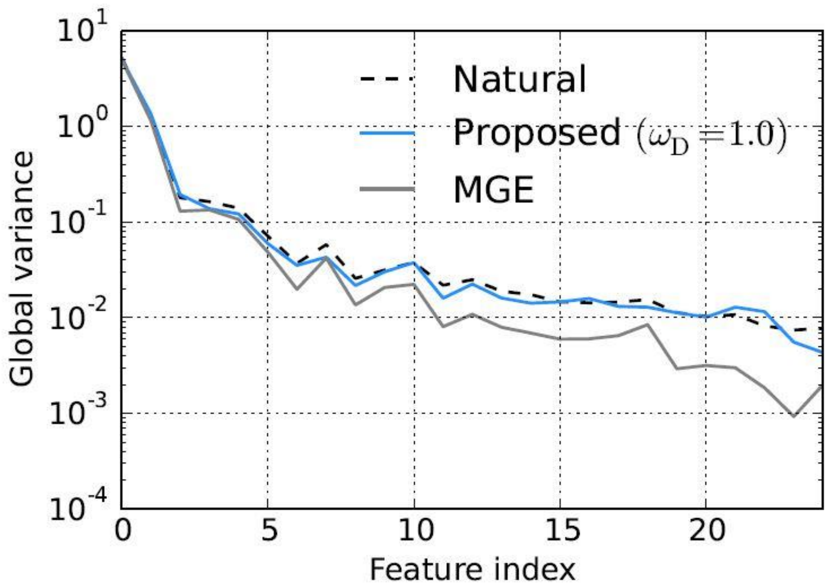
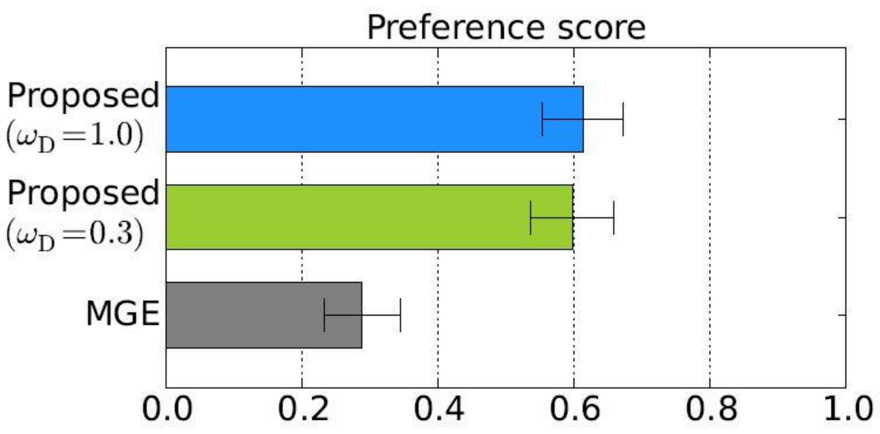


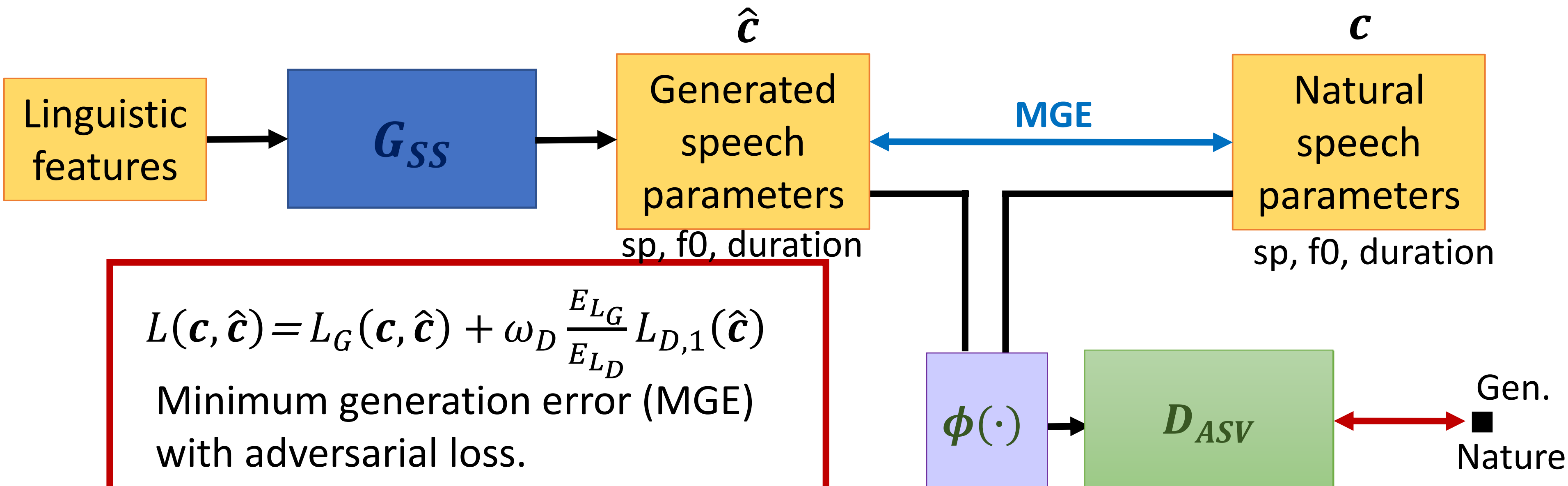
Fig. 9: Scores of speech quality.



1. The proposed algorithm generates MCCs similar to the natural ones.
2. The proposed algorithm outperforms conventional MGE training.

Speech Synthesis

- Speech synthesis with GAN (SS-GAN) [Saito et al., TASLP 2018]



$$L(\boldsymbol{c},\hat{\boldsymbol{c}}) = L_G(\boldsymbol{c},\hat{\boldsymbol{c}}) + \omega_D \frac{E_{L_G}}{E_{L_D}} L_{D,1}(\hat{\boldsymbol{c}})$$

Minimum generation error (MGE) with adversarial loss.

$$L_D(\boldsymbol{c},\hat{\boldsymbol{c}}) = L_{D,1}(\boldsymbol{c}) + L_{D,0}(\hat{\boldsymbol{c}})$$

$$L_{D,1}(\boldsymbol{c}) = -\frac{1}{T}\sum_{t=1}^{T} \log(D(\boldsymbol{c}_t))$$ ...NAT

$$L_{D,0}(\hat{\boldsymbol{c}}) = -\frac{1}{T}\sum_{t=1}^{T} \log(1 - D(\hat{\boldsymbol{c}}_t))$$ ...SYN

Speech Synthesis (SS-GAN)

- Subjective evaluations

Fig. 10: Scores of speech quality (sp).

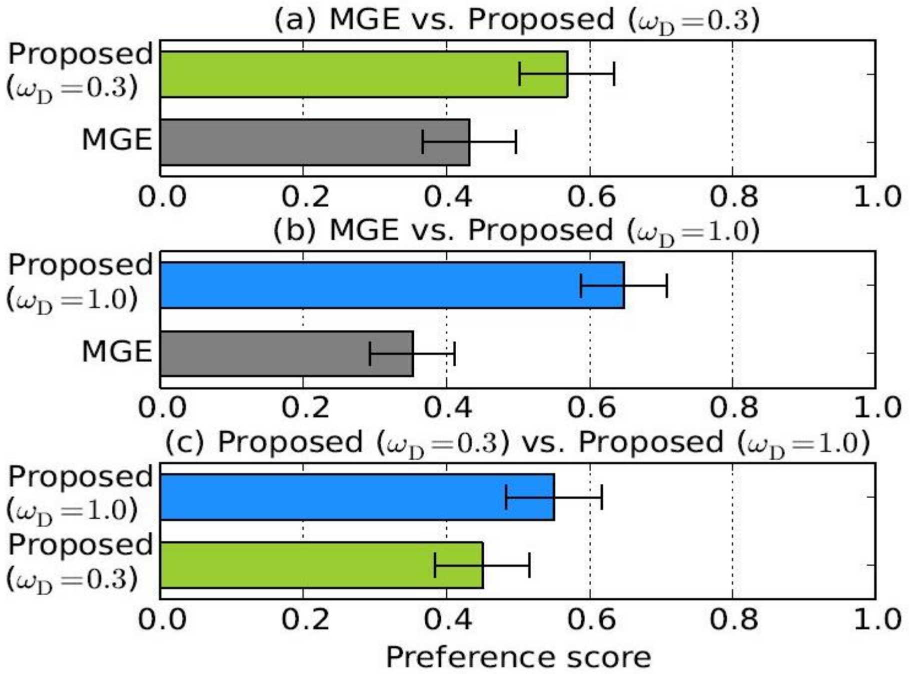
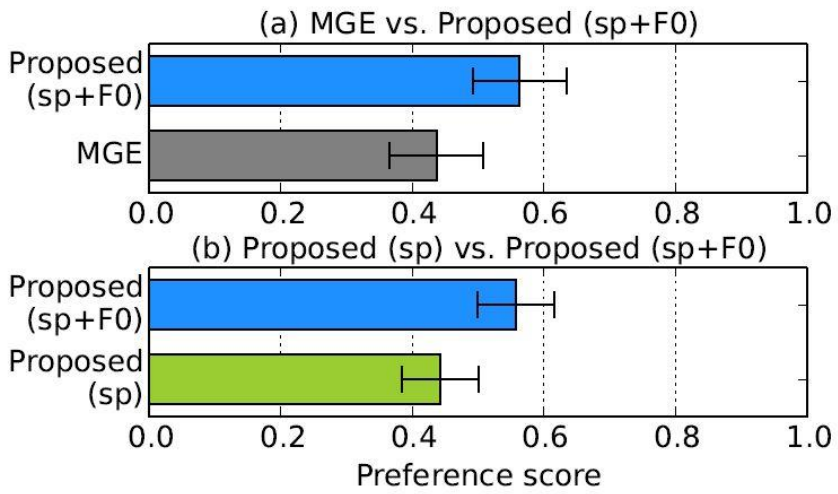


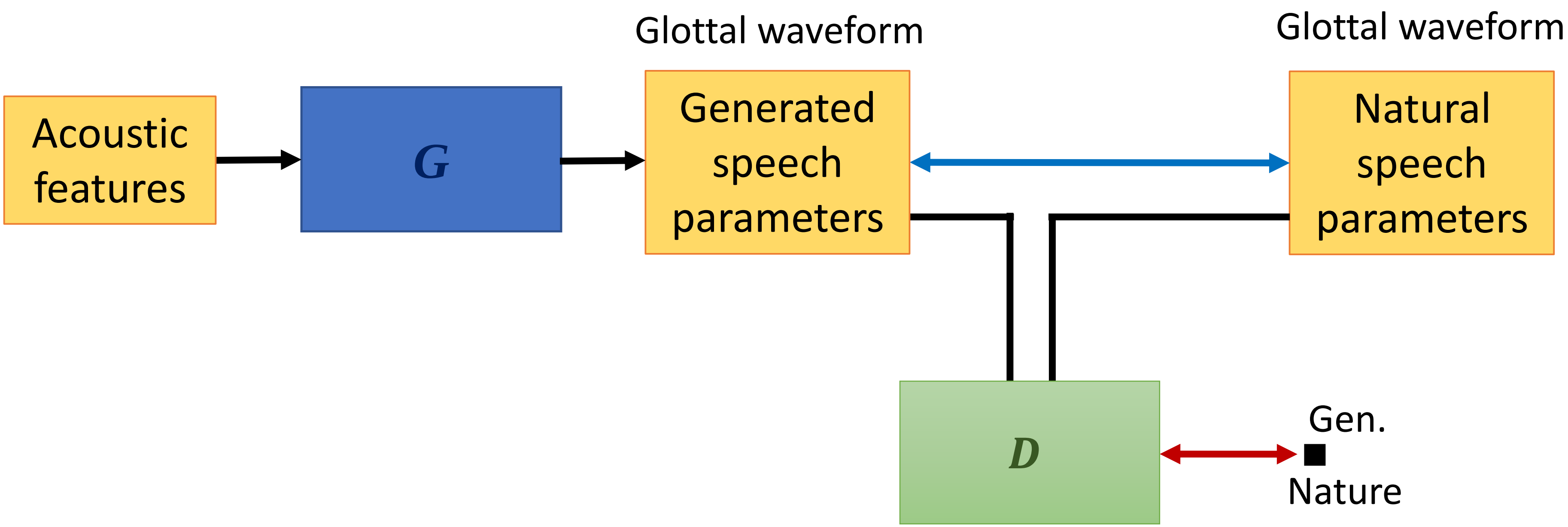
Fig. 11: Scores of speech quality (sp and F0).



The proposed algorithm works for both spectral parameters and F0.

Speech Synthesis

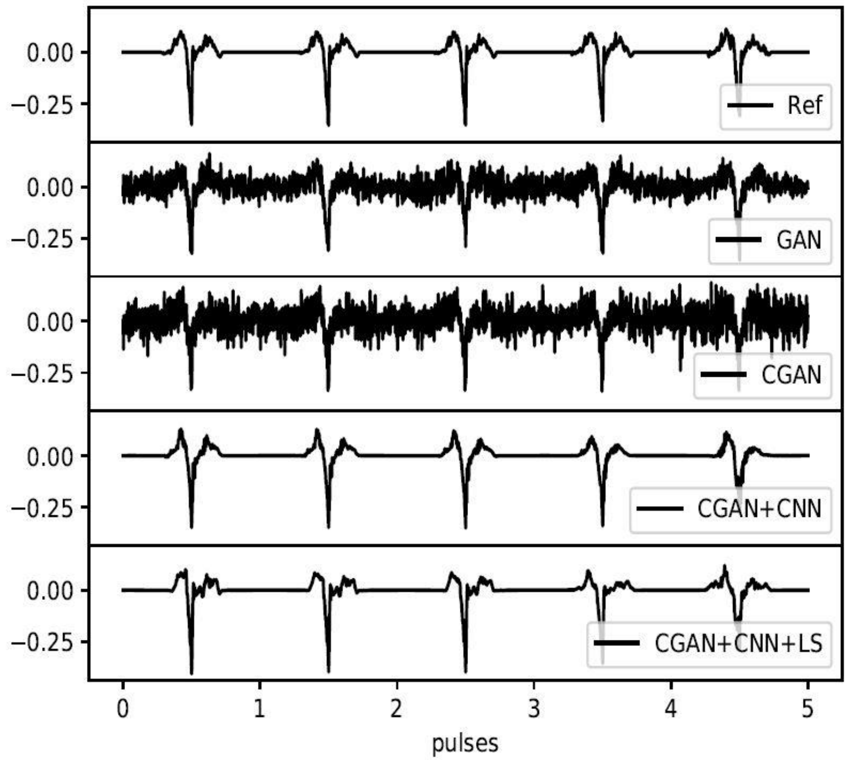
- Speech synthesis with GAN glottal waveform model (GlottGAN) [Bollepalli et al., Interspeech 2017]



Speech Synthesis (GlottGAN)

- Objective evaluations

Fig. 12: Glottal pulses generated by GANs.



G, D: DNN

G, D: conditional DNN

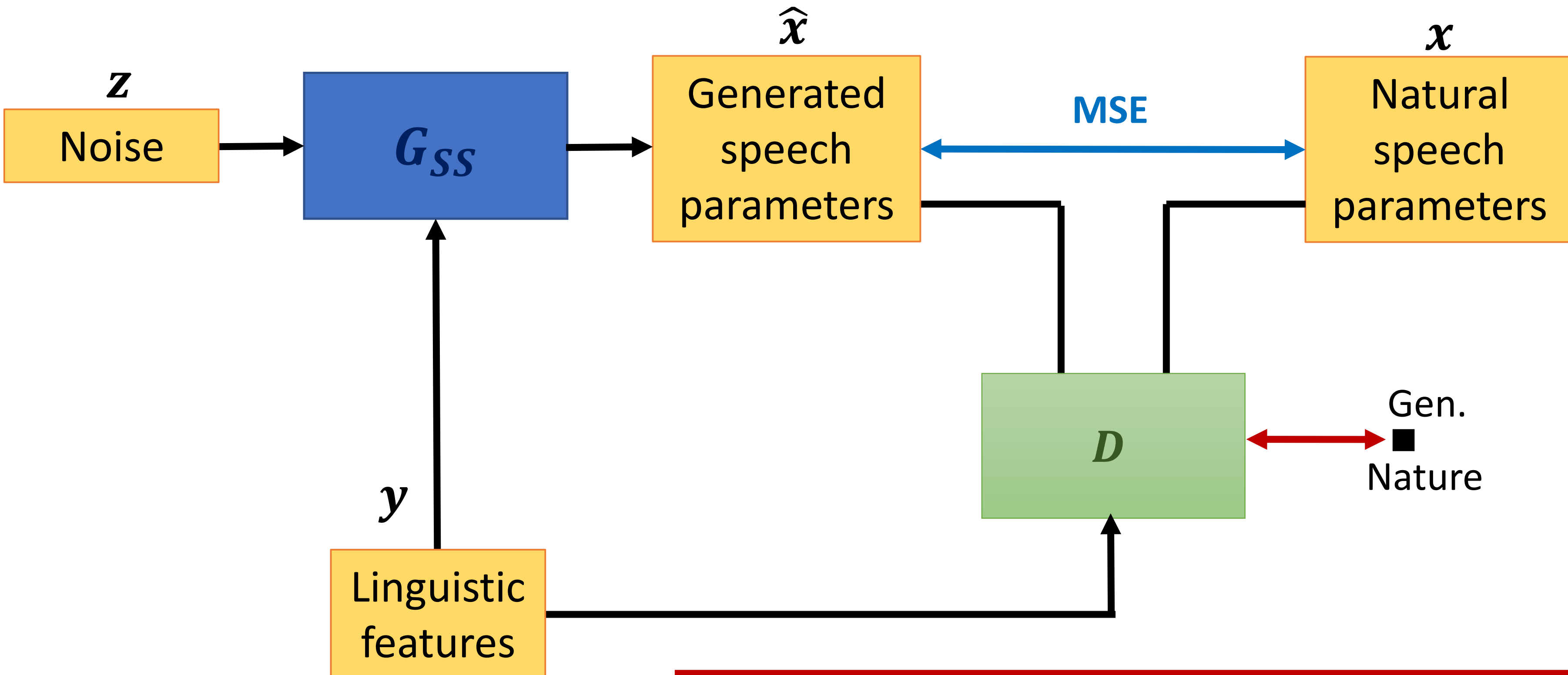
G, D: Deep CNN

G, D: Deep CNN + LS loss

The proposed GAN-based approach can generate glottal waveforms similar to the natural ones.

Speech Synthesis

- Speech synthesis with GAN & multi-task learning (SS-GAN-MTL) [Yang et al., ASRU 2017]

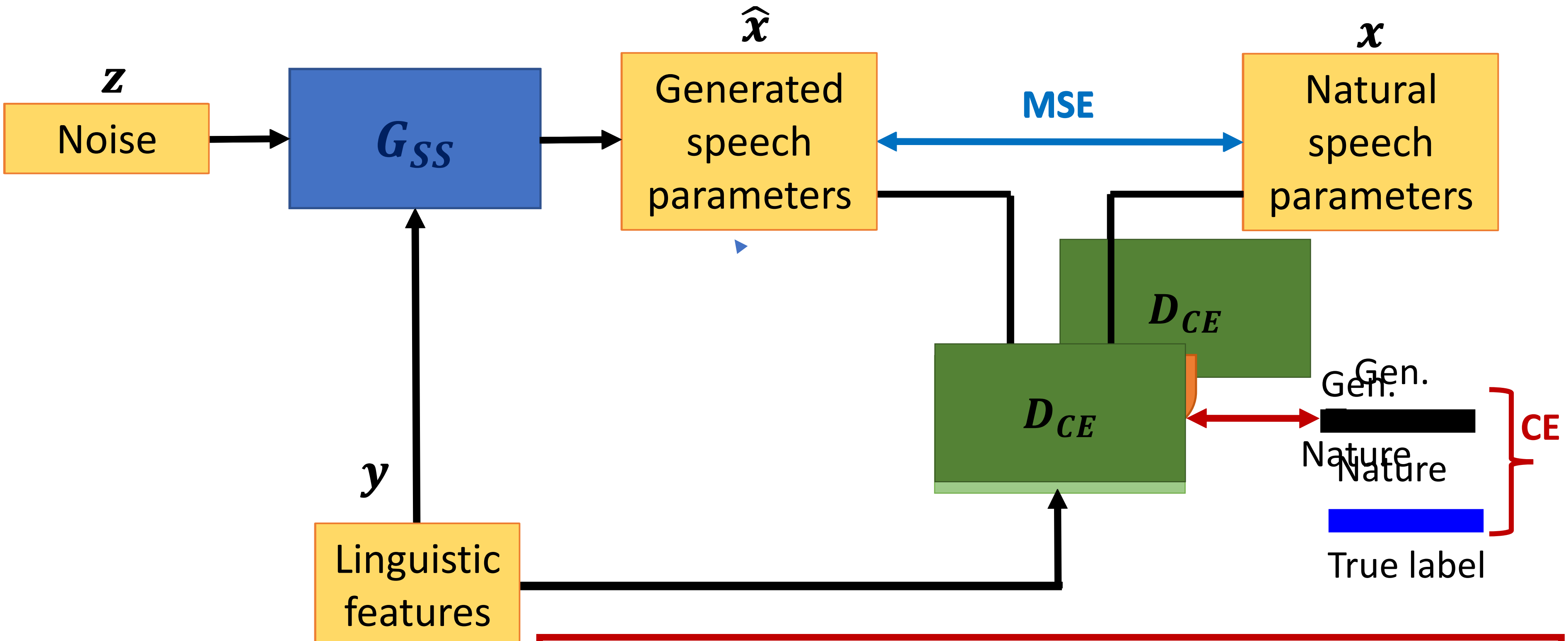


$$V_{GAN}(G,D) = E_{\boldsymbol{x}\sim p_{data}(x)}[logD(\boldsymbol{x}|\boldsymbol{y})] + E_{\boldsymbol{z}\sim p_{z},}\left[\log(1 - D\big(G(\boldsymbol{z}|\boldsymbol{y})\big)|\boldsymbol{y}\right]$$

$$V_{L2}(G) = E_{\boldsymbol{z}\sim p_{z},}[G(\boldsymbol{z}|\boldsymbol{y}) - \boldsymbol{x}]^2$$

Speech Synthesis (SS-GAN-MTL)

- Speech synthesis with GAN & multi-task learning (SS-GAN-MTL) [Yang et al., ASRU 2017]



$$V_{GAN}(G,D) = E_{x\sim p_{data}(x)}[\log D_{CE}(\boldsymbol{x}|\boldsymbol{y}, label)] + E_{z\sim p_{z},}\left[\log(1 - D_{CE}\big(G(\boldsymbol{z}|\boldsymbol{y})\big)|\boldsymbol{y}, label\right]$$

$$V_{L2}(G) = E_{z\sim p_{z},}[G(\boldsymbol{z}|\boldsymbol{y}) - \boldsymbol{x}]^2$$

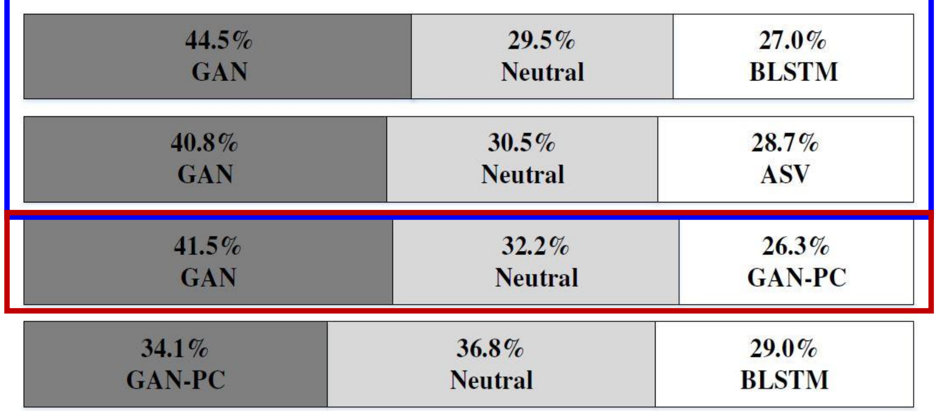
Speech Synthesis (SS-GAN-MTL)

- Objective and subjective evaluations

Table 11: Objective evaluation results.

Methods	MCD (dB)	F_0 RMSE (Hz)	V/UV (%)
BLSTM	4.624	18.544	6.447
ASV [16]	4.670	18.871	6.562
GAN	4.633	18.678	6.492
GAN-PC	4.628	18.616	6.464

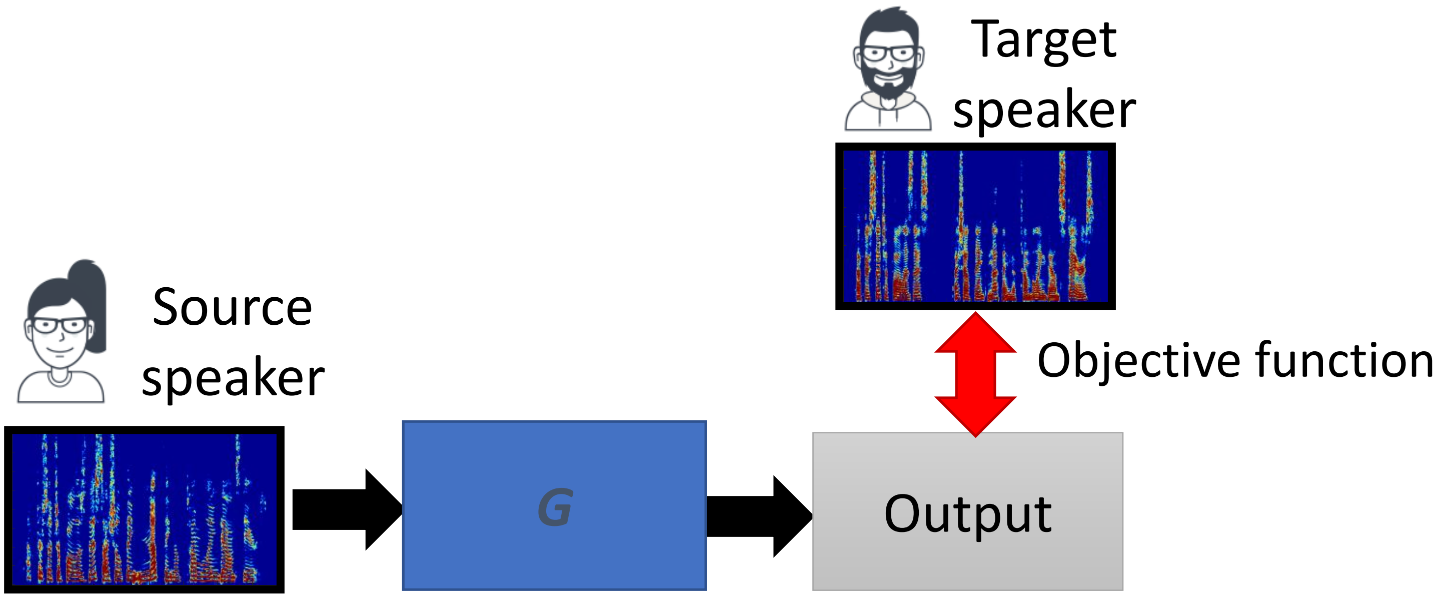
Fig. 13: The preference score (%).



1. From objective evaluations, no remarkable difference is observed.
2. From subjective evaluations, GAN outperforms BLSTM and ASV, while GAN-PC underperforms GAN.

Voice Conversion

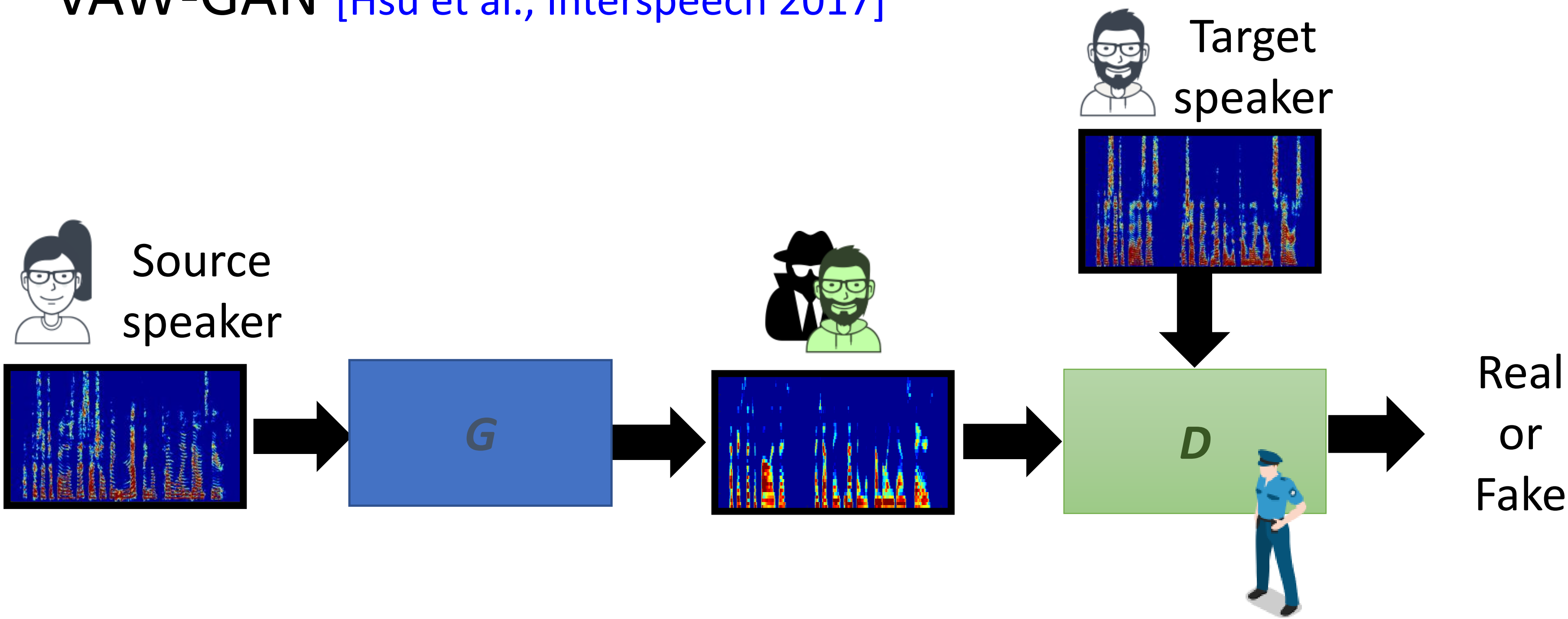
- Convert (transform) speech from source to target



➢ Conventional VC approaches include Gaussian mixture model (GMM) [Toda et al., TASLP 2007], non-negative matrix factorization (NMF) [Wu et al., TASLP 2014; Fu et al., TBME 2017], locally linear embedding (LLE) [Wu et al., Interspeech 2016], restricted Boltzmann machine (RBM) [Chen et al., TASLP 2014], feed forward NN [Desai et al., TASLP 2010], recurrent NN (RNN) [Nakashika et al., Interspeech 2014].

Voice Conversion

- VAW-GAN [Hsu et al., Interspeech 2017]



➢ Conventional MMSE approaches often encounter the "over-smoothing" issue.
➢ GAN is used a new objective function to estimate ***G***.
➢ The goal is to increase the naturalness, clarity, similarity of converted speech.

$$\boxed{V(G,D) = V_{GAN}(G,D) + \lambda V_{VAE}(\boldsymbol{x}|\boldsymbol{y})}$$

Voice Conversion (VAW-GAN)

- Objective and subjective evaluations

Fig. 14: The spectral envelopes.

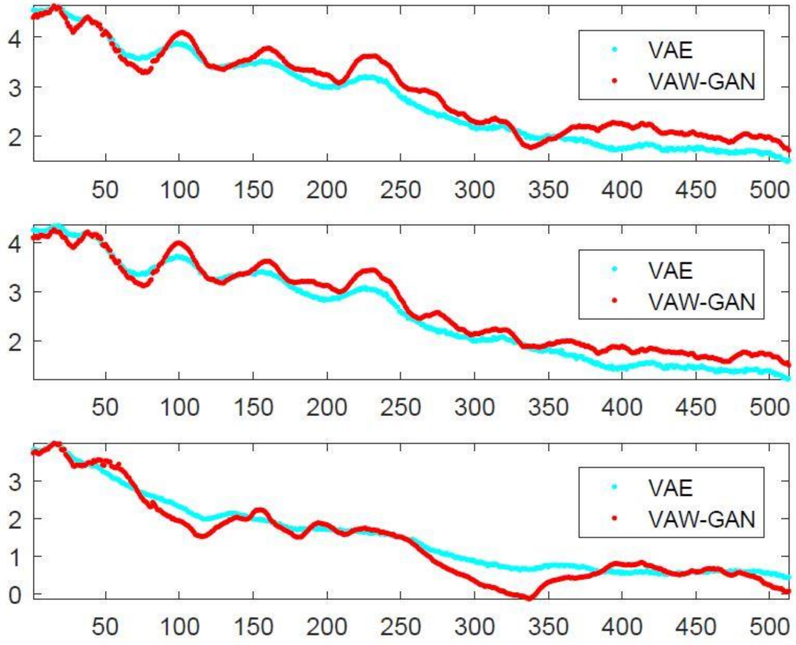
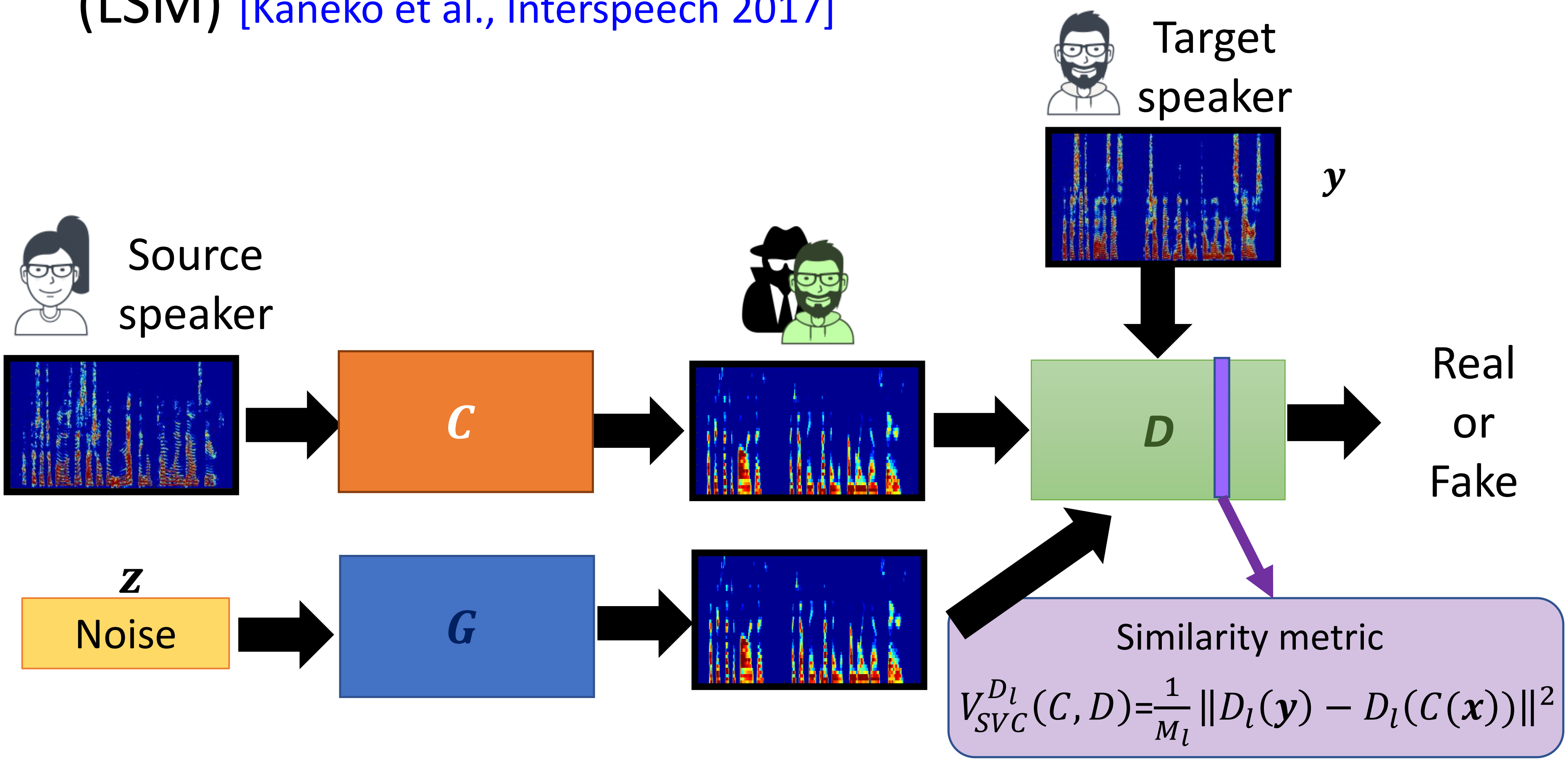


Fig. 15: MOS on naturalness.

VAW-GAN outperforms VAE in terms of objective and subjective evaluations with generating more structured speech.

Voice Conversion

- Sequence-to-sequence VC with learned similarity metric (LSM) [Kaneko et al., Interspeech 2017]

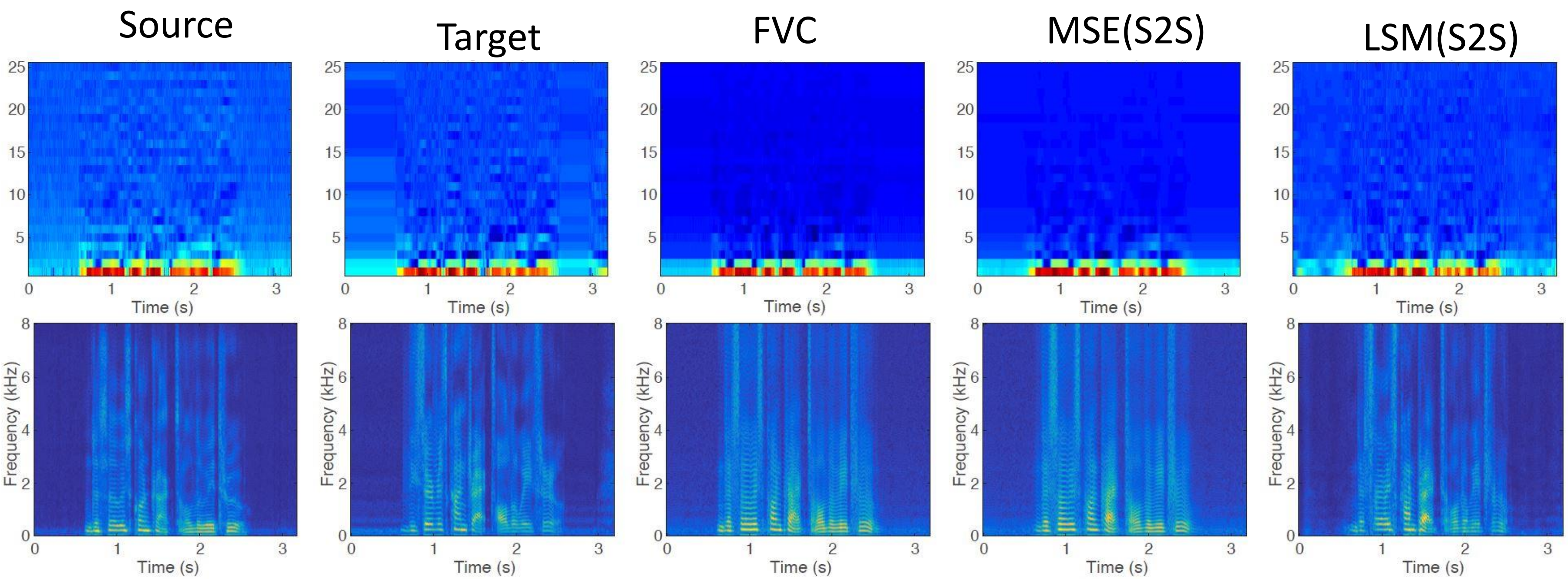


$$\boxed{V(C,G,D) = V_{SVC}^{D_l}(C,D) + V_{GAN}(C,G,D)}$$

Voice Conversion (LSM)

- Spectrogram analysis

Fig. 16: Comparison of MCCs (upper) and STFT spectrograms (lower).



The spectral textures of LSM are more similar to the target ones.

Voice Conversion (LSM)

- Subjective evaluations

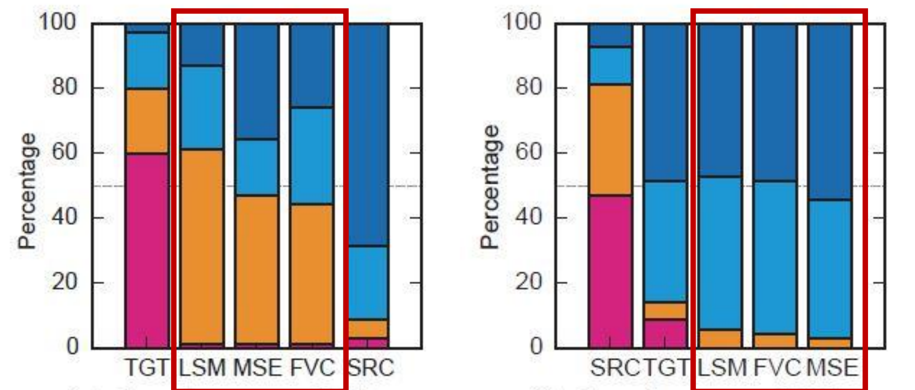
Table 12: Preference scores for naturalness.

	Former	Latter	Neutral
FVC vs. LSM	17.1 ± 6.3	**72.9** ± 7.5	10.0 ± 5.0
MSE vs. LSM	10.0 ± 5.0	**84.3** ± 6.1	5.7 ± 3.9

Table 12: Preference scores for clarity.

	Former	Latter	Neutral
FVC vs. LSM	32.9 ± 7.9	**54.3** ± 8.4	12.9 ± 5.6
MSE vs. LSM	27.1 ± 7.5	**65.0** ± 8.0	7.9 ± 4.5

Fig. 17: Similarity of TGT and SRC with VCs.



(a) Similarity to TGT *(b) Similarity to SRC*

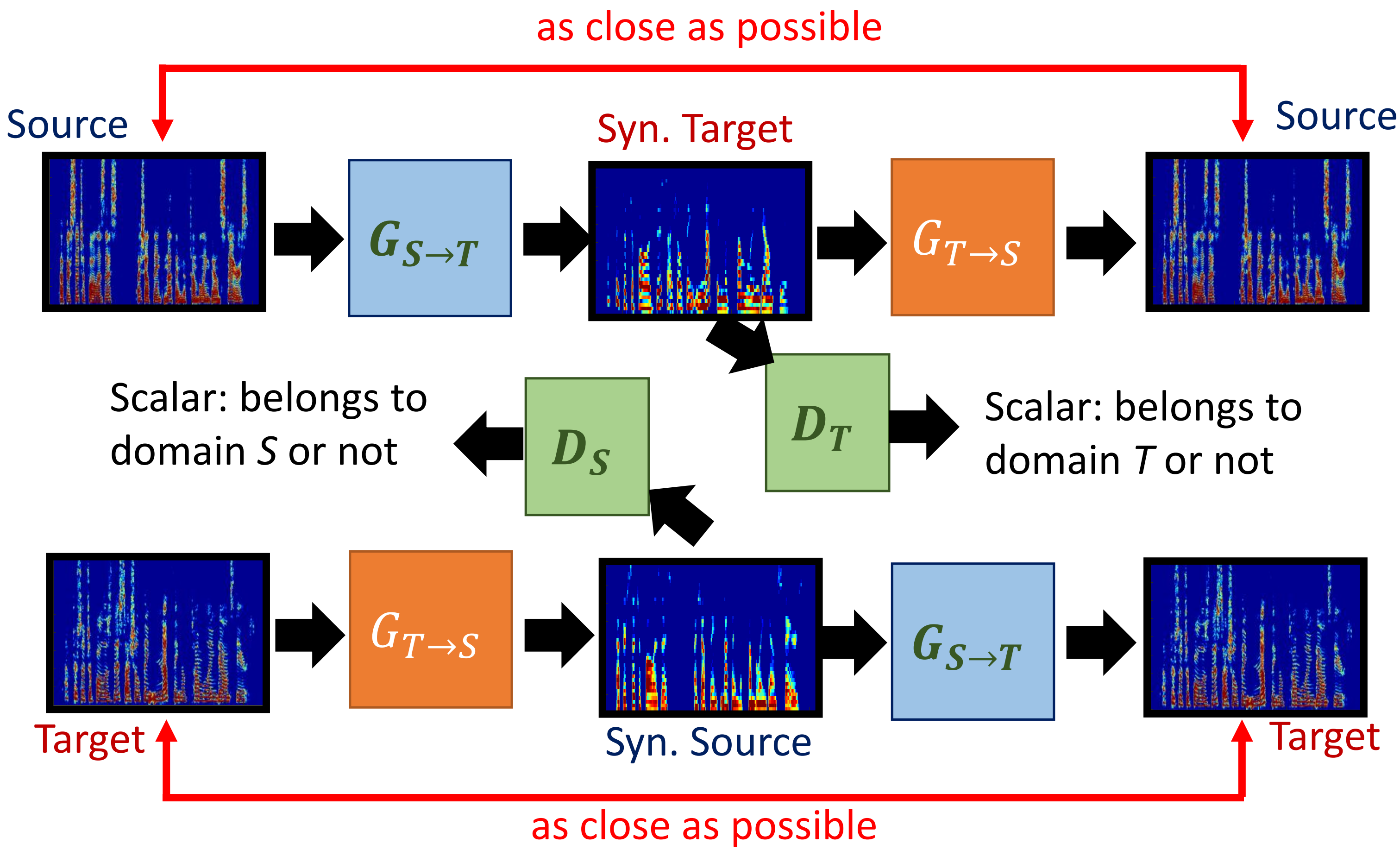
Target speaker

Source speaker

LSM outperforms FVC and MSE in terms of subjective evaluations.

Voice Conversion

- CycleGAN-VC [Kaneko et al., arXiv 2017]



$$
\boxed{\begin{aligned} V_{Full} = {} & V_{GAN}(G_{X\to Y}, D_Y) + V_{GAN}(G_{X\to Y}, D_Y) \\ & + \lambda\, V_{Cyc}(G_{X\to Y}, G_{Y\to X}) \end{aligned}}
$$

Voice Conversion (CycleGAN-VC)

- Subjective evaluations

Fig. 18: MOS for naturalness.

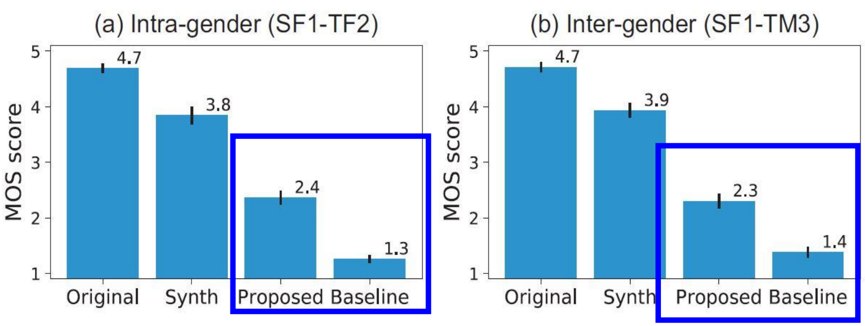
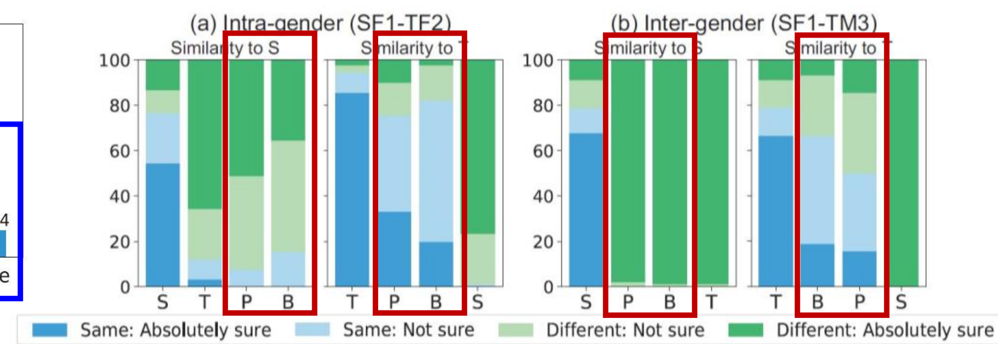


Fig. 19: Similarity of to source and to target speakers. S: Source; T:Target; P: Proposed; B:Baseline



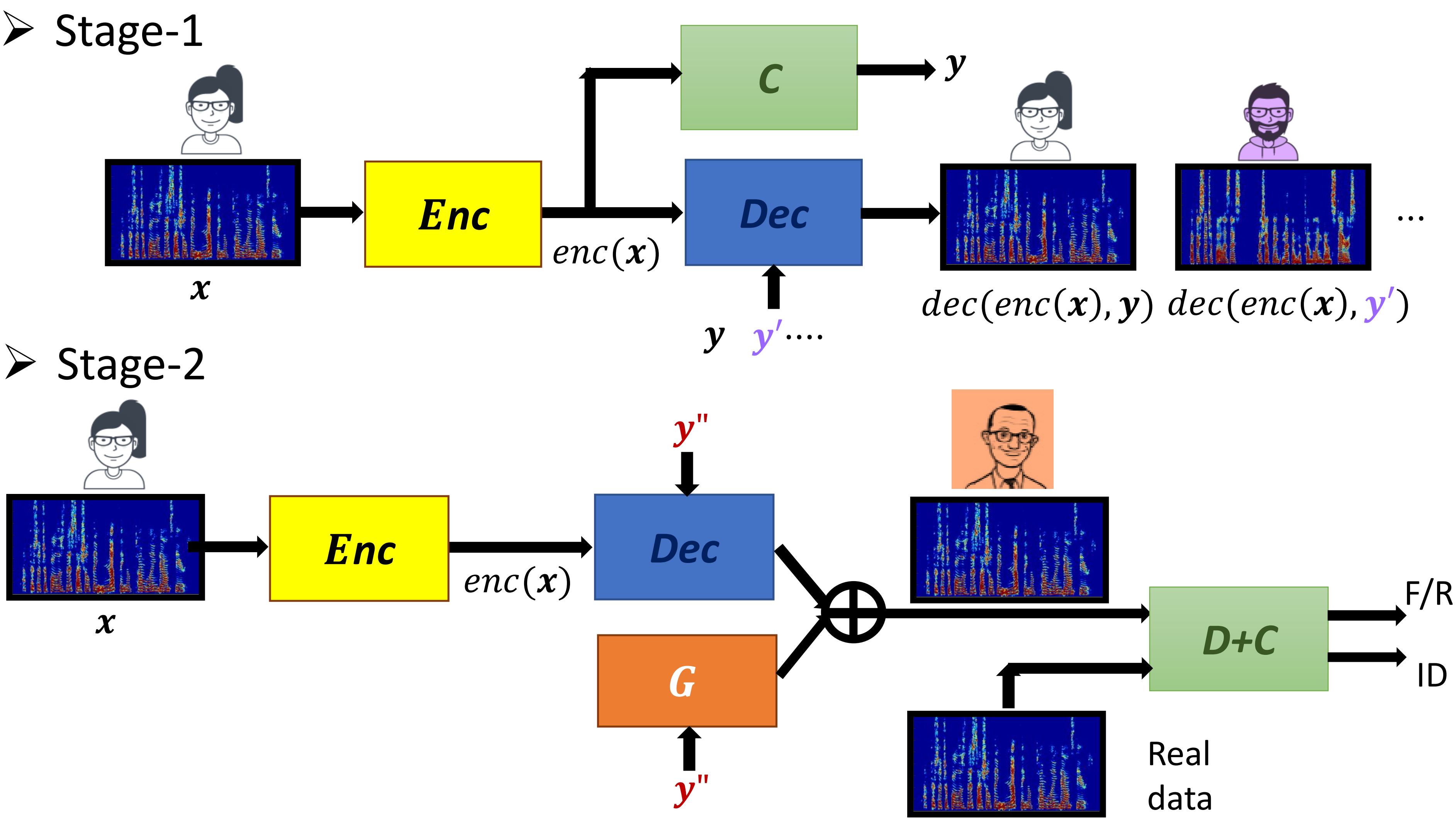
Target speaker

Source speaker

1. The proposed method uses **non-parallel** data.
2. For naturalness, the proposed method outperforms baseline.
3. For similarity, the proposed method is comparable to the baseline.

Voice Conversion

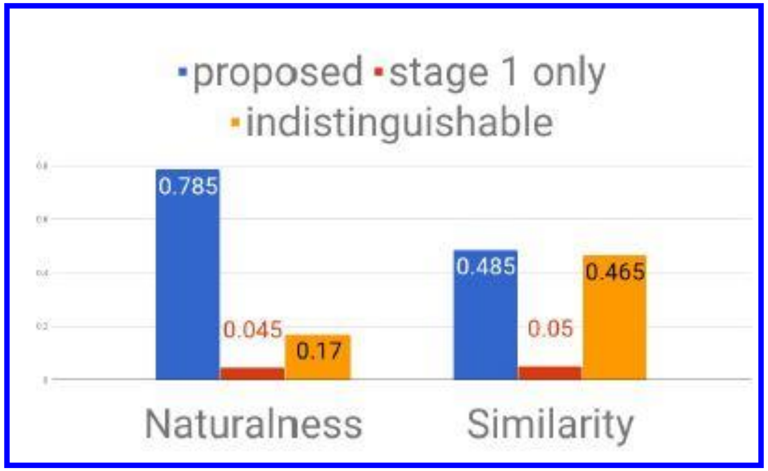
- Multi-target VC [Chou et al., arxiv 2018]

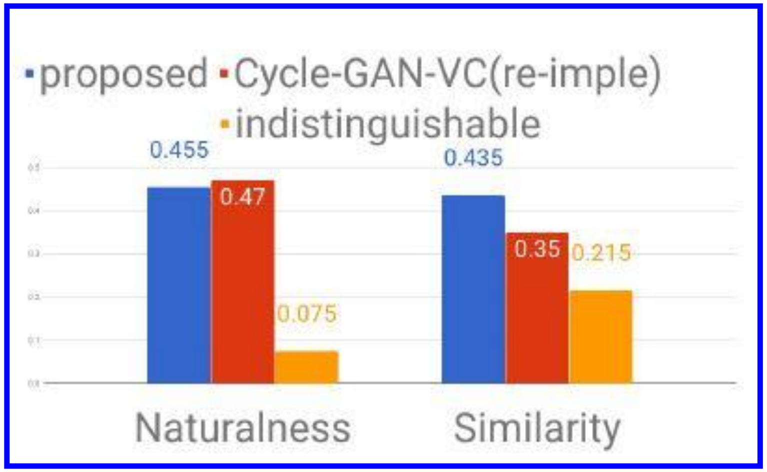


Voice Conversion (Multi-target VC)

- Subjective evaluations

Fig. 20: Preference test results





1. The proposed method uses **non-parallel** data.
2. The multi-target VC approach outperforms one-stage only.
3. The multi-target VC approach is comparable to Cycle-GAN-VC in terms of the naturalness and the similarity.

Outline of Part II

Speech Signal Generation

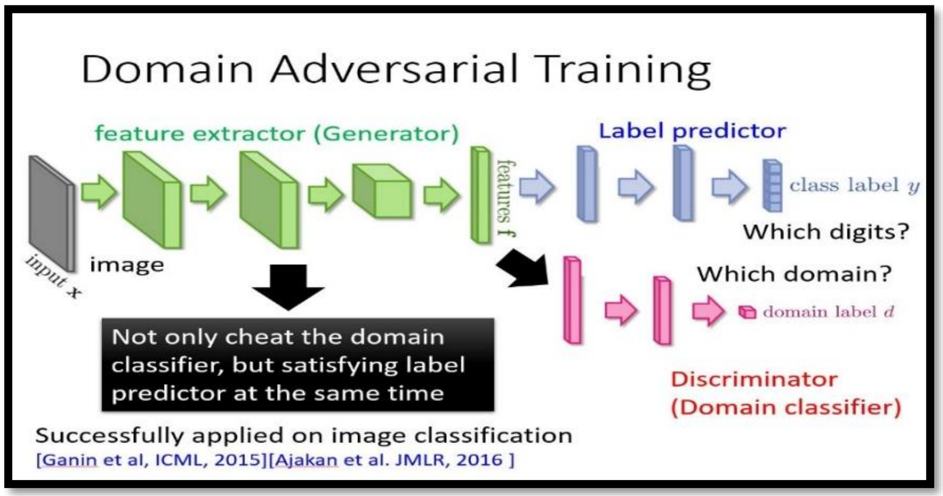
- Speech enhancement
- Postfilter, speech synthesis, voice conversion

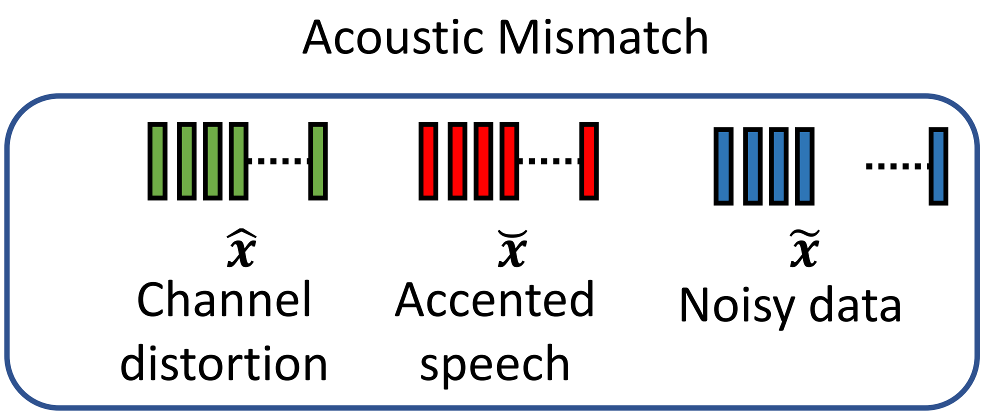
Speech Signal Recognition

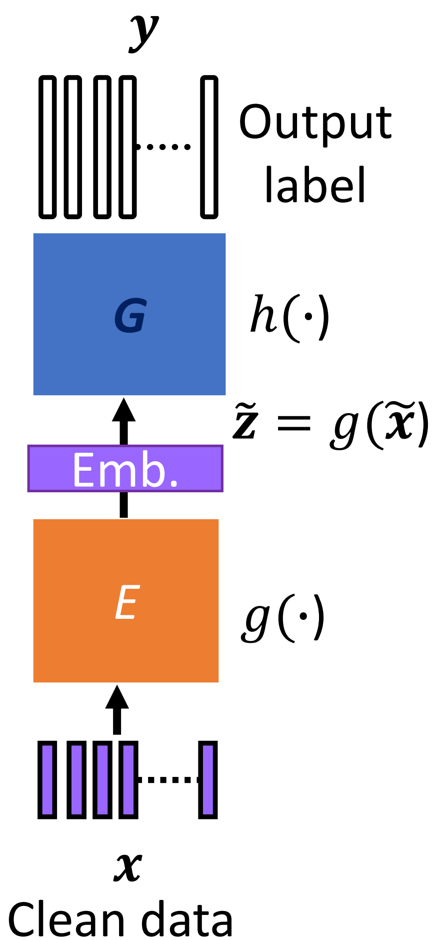
- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion

Speech, Speaker, Emotion Recognition and Lip-reading (Classification Task)

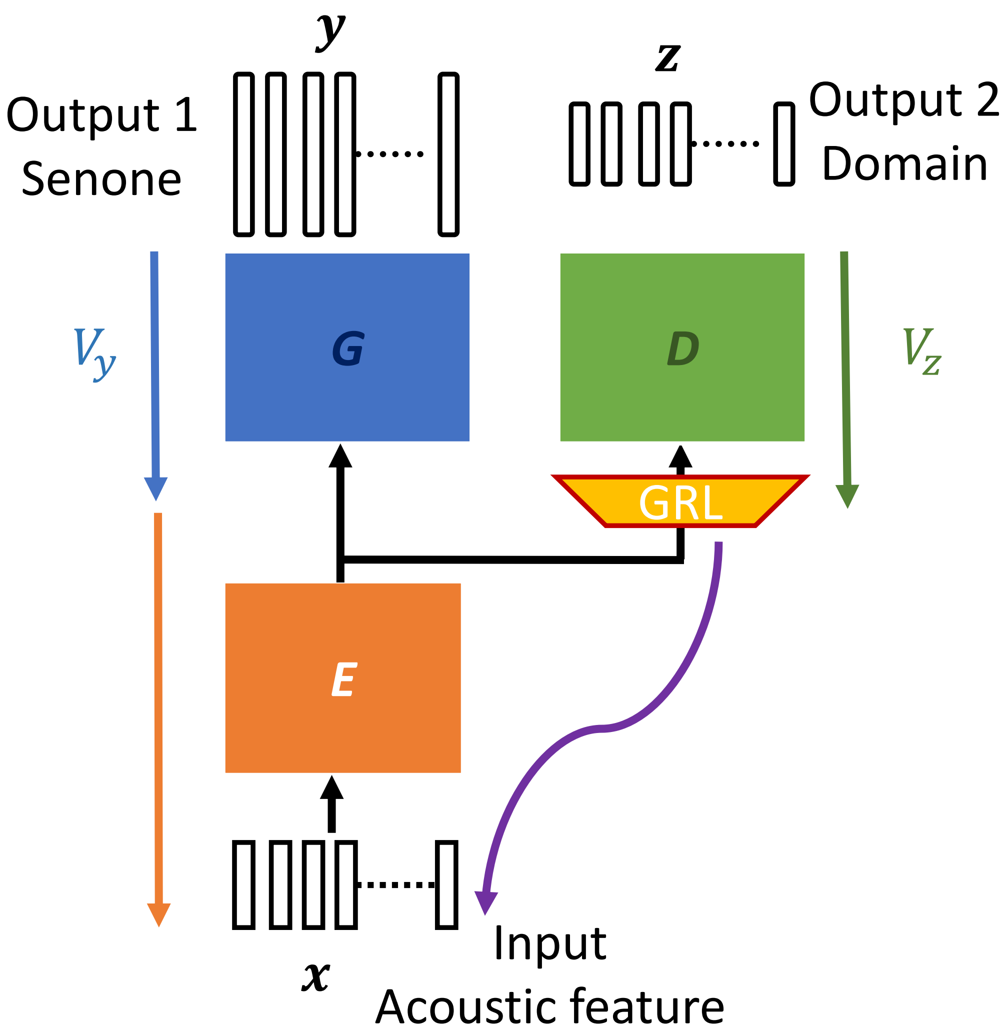






Speech Recognition

- Adversarial multi-task learning (AMT)
  [Shinohara Interspeech 2016]



Objective function

$$V_y = -\sum_i \log P(y_i | x_i; \theta_E, \theta_G)$$

$$V_z = -\sum_i \log P(z_i | x_i; \theta_E, \theta_D)$$

Model update

$$\theta_G \leftarrow \theta_G - \epsilon \frac{\partial V_y}{\partial \theta_G}$$

Max classification accuracy

$$\theta_D \leftarrow \theta_D - \epsilon \frac{\partial V_z}{\partial \theta_D}$$

Max domain accuracy

$$\theta_E \leftarrow \theta_E - \epsilon \left(\frac{\partial V_y}{\partial \theta_G} \right) + \alpha \frac{\partial V_z}{\partial \theta_G}$$

Max classification accuracy
and Min domain accuracy

Speech Recognition (AMT)

- ASR results in known (k) and unknown (unk) noisy conditions

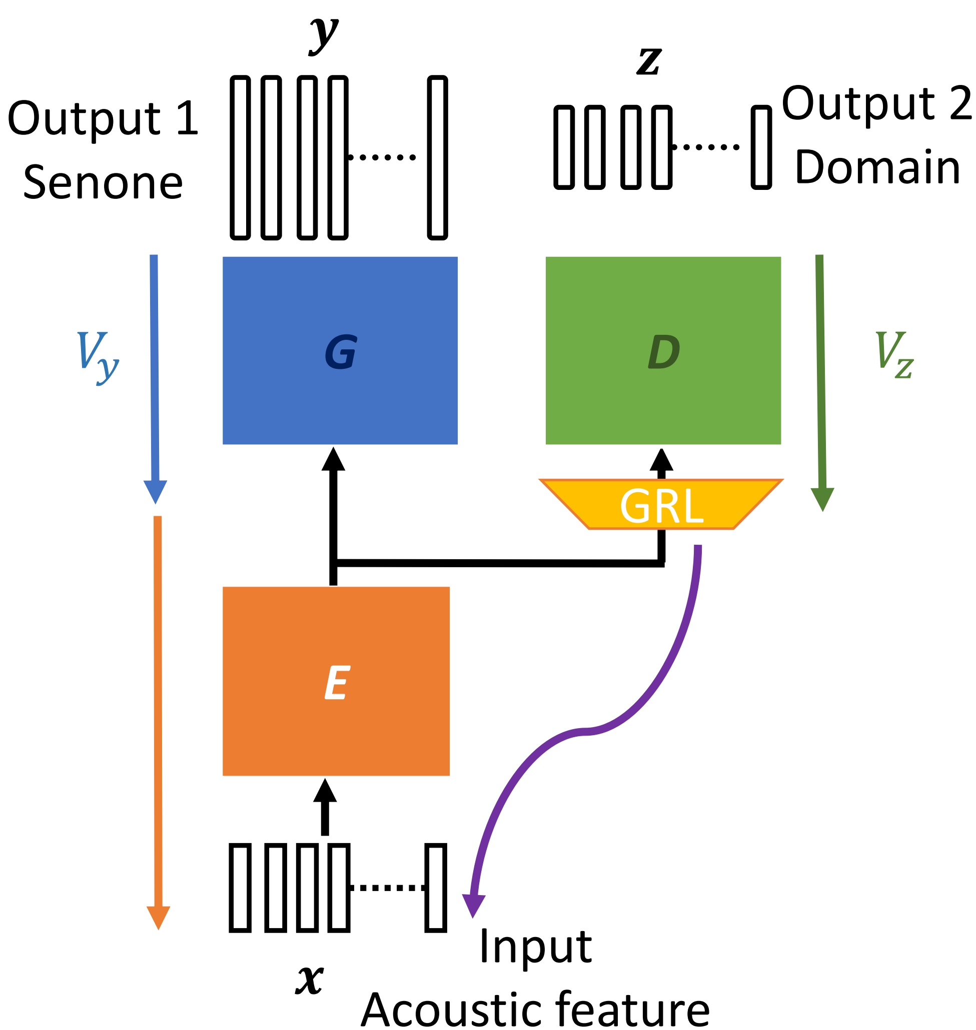
Table 13: WER of DNNs with single-task learning (ST) and AMT.

	noise	ST	AMT	RERR
k	car 2000cc	5.83	5.56	4.63
k	exhib. booth	6.80	6.66	2.06
k	station	7.89	7.76	1.65
k	crossing	6.96	6.65	4.45
unk	car 1500cc	5.58	5.46	2.15
unk	exhib. aisle	7.71	6.93	10.12
unk	factory	12.17	12.92	-6.16
unk	highway	9.73	9.52	2.16
unk	crowd	6.72	6.40	4.76
unk	server room	8.54	7.76	9.13
unk	air cond.	6.96	6.98	-0.29
unk	elev. hall	9.23	9.60	-4.01
-	average	7.84	7.68	2.04

The AMT-DNN outperforms ST-DNN with yielding lower WERs.

Speech Recognition

- Domain adversarial training for accented ASR (DAT) [Sun et al., ICASSP2018]



Objective function

$$V_y = -\sum_i \log P(y_i | x_i; \theta_E, \theta_G)$$

$$V_z = -\sum_i \log P(z_i | x_i; \theta_E, \theta_D)$$

Model update

$$\theta_G \leftarrow \theta_G - \epsilon \frac{\partial V_y}{\partial \theta_G}$$

Max classification accuracy

$$\theta_D \leftarrow \theta_D - \epsilon \frac{\partial V_z}{\partial \theta_D}$$

Max domain accuracy

$$\theta_E \leftarrow \theta_E - \epsilon \left(\frac{\partial V_y}{\partial \theta_G} \right) + \alpha \frac{\partial V_z}{\partial \theta_G}$$

Max classification accuracy and Min domain accuracy

Speech Recognition (DAT)

- ASR results on accented speech

Table 14: WER of the baseline and adapted model.

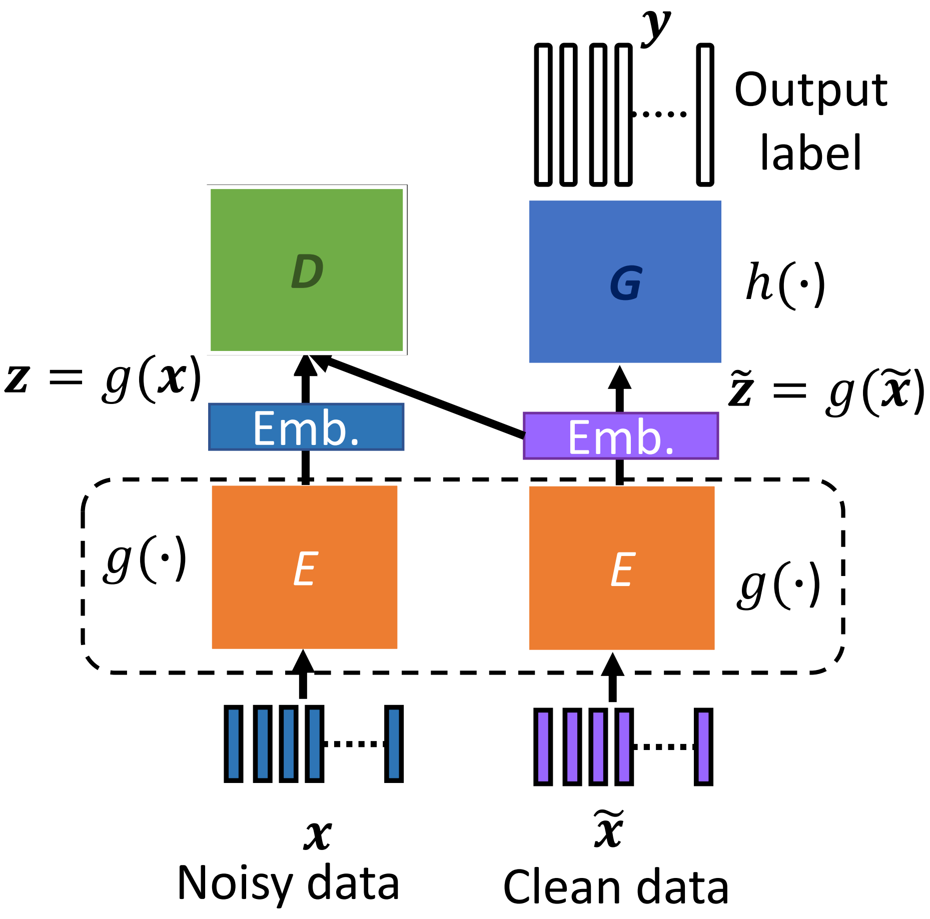
training data	λ	test							
		STD	FJ	JS	JX	SC	GD	HN	Avg.
STD	-	15.55	23.58	15.75	14.08	15.62	15.32	19.34	17.28
STD + (600hrs with trans)	-	14.22	14.84	9.41	8.68	9.13	9.62	11.89	10.60
STD + (600hrs no trans)	0.03	15.37	22.96	14.48	13.79	15.35	14.86	18.24	16.61

STD: standard speech

1. With labeled transcriptions, ASR performance notably improves.
2. DAT is effective in learning features invariant to domain differences with and without labeled transcriptions.

Speech Recognition

- Robust ASR using GAN enhancer (GAN-Enhancer)
  [Sriram et al., arXiv 2017]



Cross entropy with L1 Enhancer:

$$H(h(\tilde{\mathbf{z}}),\mathbf{y}) + \lambda \frac{\|\boldsymbol{z} - \tilde{\mathbf{z}}\|_1}{\|\boldsymbol{z}\|_1 + \|\tilde{\mathbf{z}}\|_1 + \epsilon}$$

Cross entropy with GAN Enhancer:

$$H(h(\tilde{\mathbf{z}}),\mathbf{y}) + \lambda\, V_{adv}\,(g(\boldsymbol{x})\,,\, g(\tilde{\boldsymbol{x}}))$$

Speech Recognition (GAN-Enhancer)

- ASR results on far-field speech:

Fig. 15: WER of GAN enhancer and the baseline methods.

Model	Near-Field		Far-Field	
	CER	WER	CER	WER
seq-to-seq	7.43%	21.18%	23.76%	50.84%
seq-to-seq + far-field Augmentation	7.69%	21.32%	12.47%	30.59%
seq-to-seq + L^1-Distance Penalty	7.54%	20.45%	12.00%	29.19%
seq-to-seq + GAN Enhancer	7.78%	21.07%	**11.26%**	**28.12%**

GAN Enhancer outperforms the Augmentation and L1-Enhancer approaches on far-field speech.

Outline of Part II

Speech Signal Generation

- Speech enhancement
- Postfilter, speech synthesis, voice conversion

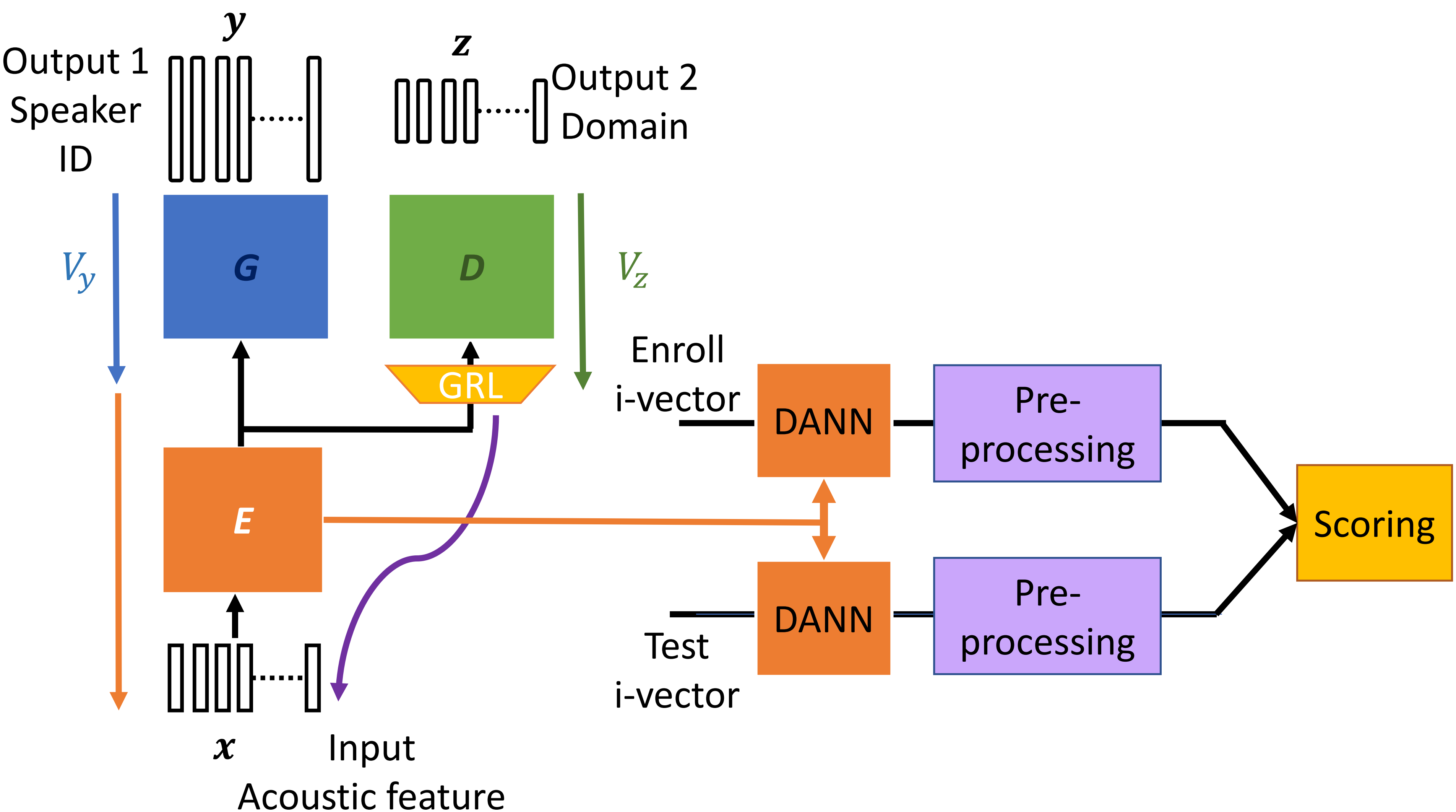
Speech Signal Recognition

- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion

Speaker Recognition

- Domain adversarial neural network (DANN)
  [Wang et al., ICASSP 2018]



Speaker Recognition (DANN)

- Recognition results of domain mismatched conditions

Table 16: Performance of DAT and the state-of-the-art methods.

Systems#	Adaptation Methods	EER%	DCF10 [21]	DCF08
1	–	9.35	0.724	0.520
2	–	5.66	0.633	0.427
3	Interpolated [6] [12]	6.55	0.652	0.454
4	IDV [9] [12]	6.15	0.676	0.476
5	DICN [11] [12]	4.99	0.623	0.416
6	DAE [22] [12]	4.81	0.610	0.398
7	AEDA [12]	4.50	0.589	0.362
8	**DAT**	**3.73**	**0.541**	**0.335**

The DAT approach outperforms other methods with achieving lowest EER and DCF scores.

Outline of Part II

Speech Signal Generation

- Speech enhancement
- Postfilter, speech synthesis, voice conversion

Speech Signal Recognition

- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

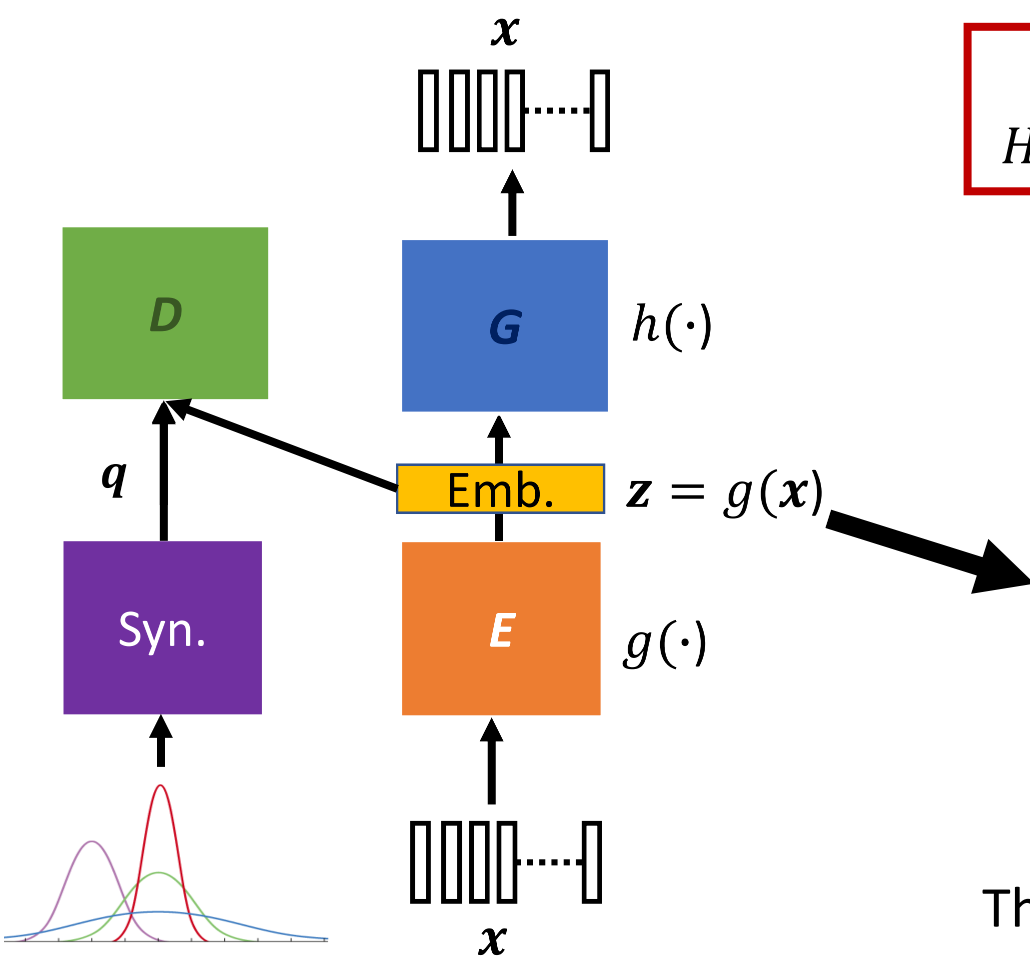
Conclusion

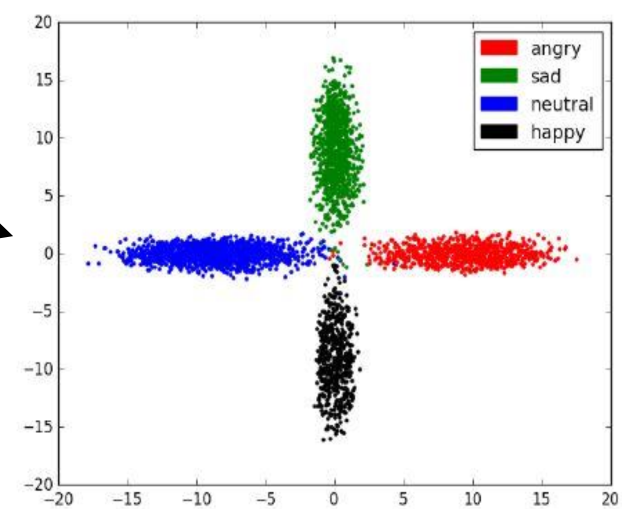
Emotion Recognition

- Adversarial AE for emotion recognition (AAE-ER)
  [Sahu et al., Interspeech 2017]

AE with GAN :

$$H(h(\boldsymbol{z}), \boldsymbol{x}) + \lambda \, V_{GAN} \, (\boldsymbol{q}, g(\boldsymbol{x}))$$





The distribution of code vectors

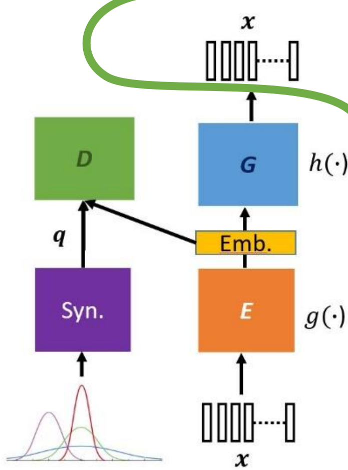
Emotion Recognition (AAE-ER)

- Recognition results of domain mismatched conditions:

Table 17: Classification results on different systems.

	OpenSmile features (1582-D)	Code vectors (2-D)	Auto-encoder (100-D)	LDA (2-D)	PCA (2-D)
UAR (%)	57.88	56.38	53.92	48.67	43.12

Table 18: Class ... n real and synthesized features.



	UAR (%)
...c	25.00
S ...ts	33.75
...only	57.88
Sy ...tapoints	**58.38**

1. AAE alone could not yield performance improvements.
2. Using synthetic data from AAE can yield higher UAR.

Outline of Part II

Speech Signal Generation

- Speech enhancement
- Postfilter, speech synthesis, voice conversion

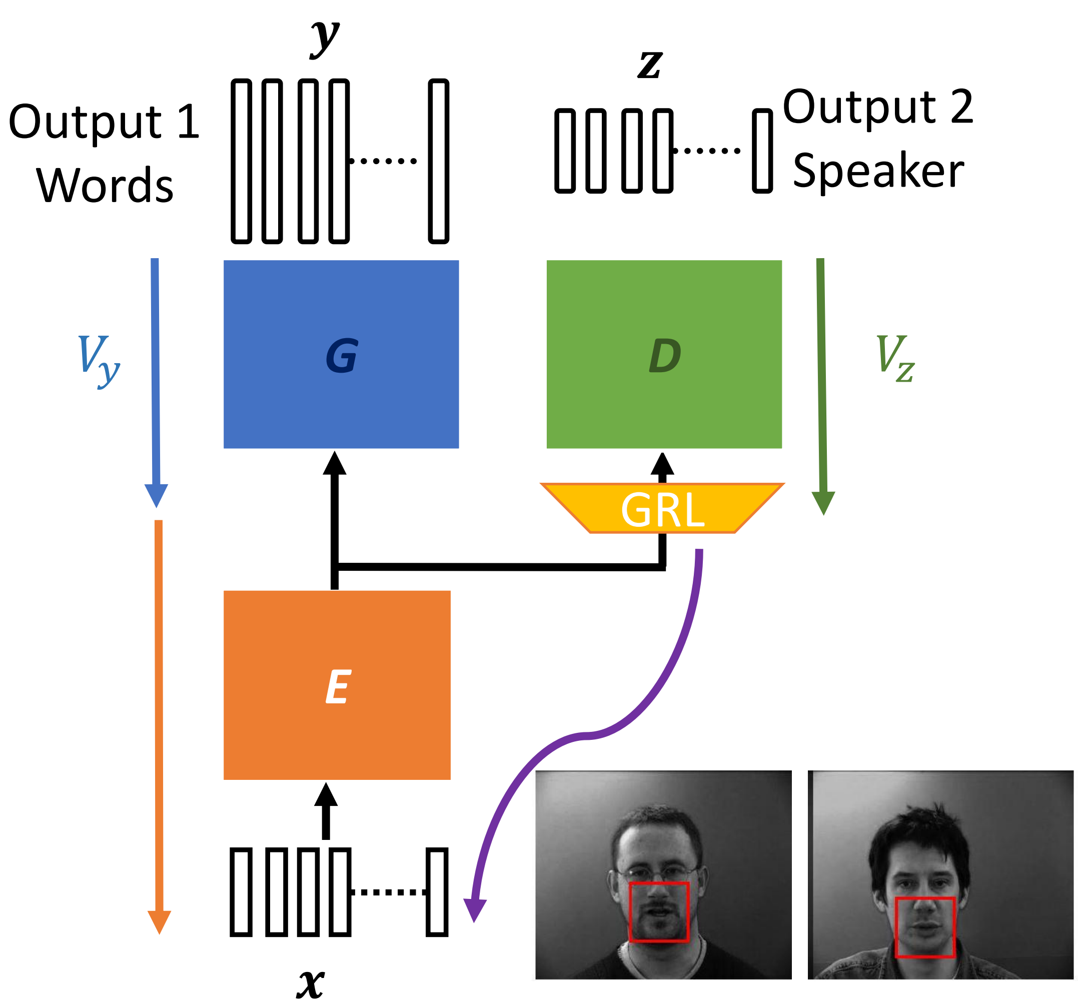
Speech Signal Recognition

- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion

Lip-reading

- Domain adversarial training for lip-reading (DAT-LR) [Wand et al., arXiv 2017]



~80% WAC

Objective function

$$V_y = -\sum_i \log P(y_i | x_i; \theta_E, \theta_G)$$

$$V_z = -\sum_i \log P(z_i | x_i; \theta_E, \theta_D)$$

Model update

$$\theta_G \leftarrow \theta_G - \epsilon \frac{\partial V_y}{\partial \theta_G}$$

Max classification accuracy

$$\theta_D \leftarrow \theta_D - \epsilon \frac{\partial V_z}{\partial \theta_D}$$

Max domain accuracy

$$\theta_E \leftarrow \theta_E - \epsilon \left(\frac{\partial V_y}{\partial \theta_G} \right) + \alpha \frac{\partial V_z}{\partial \theta_G}$$

Max classification accuracy and Min domain accuracy

Lip-reading (DAT-LR)

- Recognition results of speaker mismatched conditions

Table 19: Performance of DAT and the baseline.

Adversarial Training on	Number of training spk	Target Test acc.	Relative Improvement	p-value
None	1	18.7%	-	-
	4	39.4%	-	-
	8	46.5%	-	-
All Target Sequences	1	25.4%	35.8%	0.0030*
	4	43.6%	10.7%	0.0261*
	8	49.3%	6.0%	0.0266*
50 Target Sequences	1	24.1%	28.9%	0.0045*
	4	41.5%	5.3%	0.1367
	8	47.0%	1.1%	0.3555

The DAT approach notably enhances the recognition accuracies in different conditions.

Outline of Part II

Speech Signal Generation

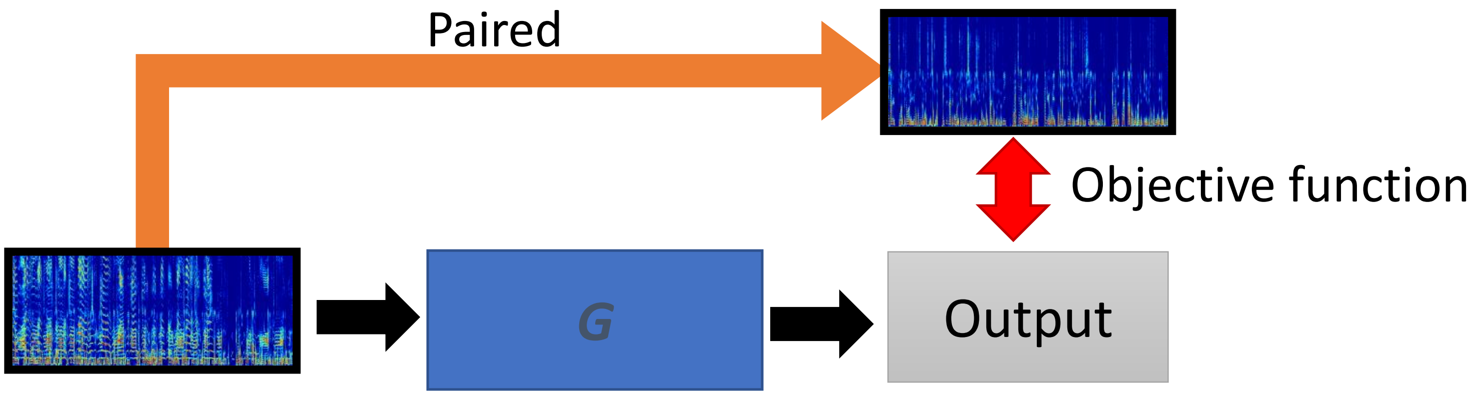
- Speech enhancement
- Postfilter, speech synthesis, voice conversion

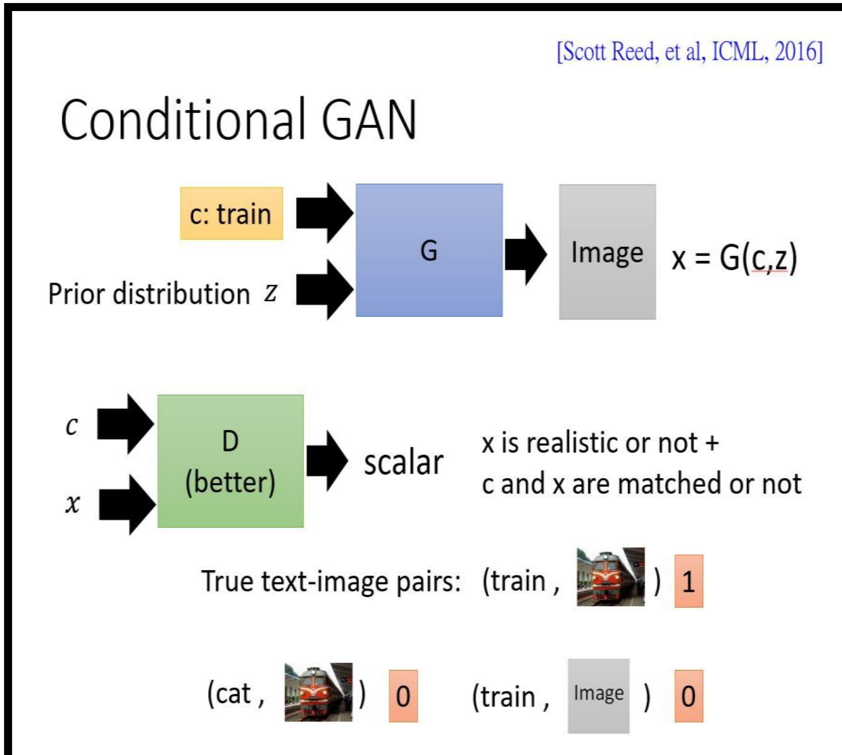
Speech Signal Recognition

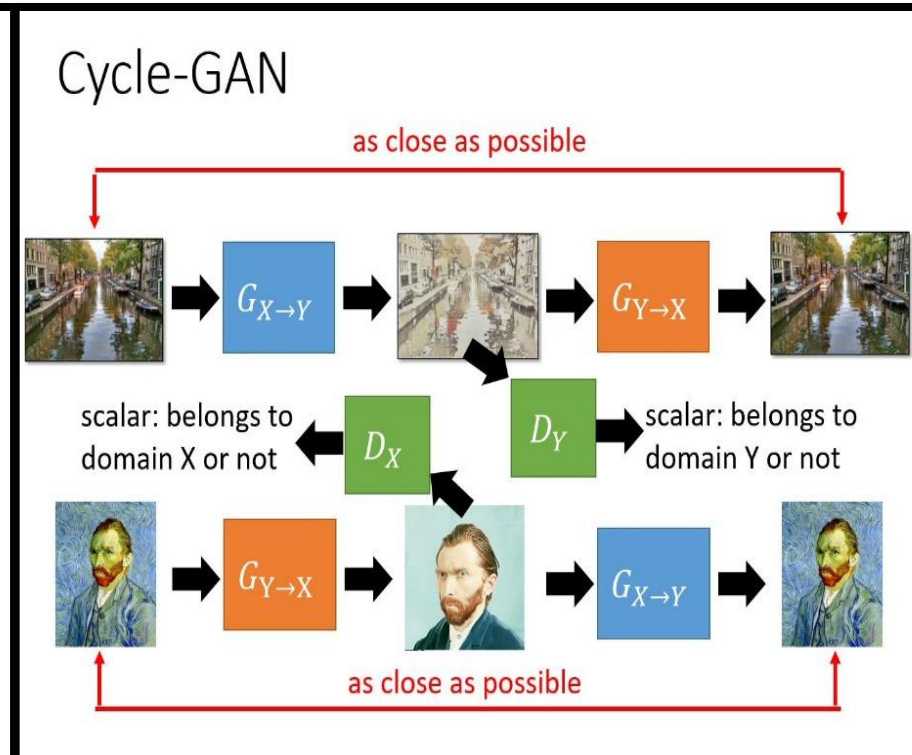
- Speech recognition
- Speaker recognition
- Speech emotion recognition
- Lip reading

Conclusion

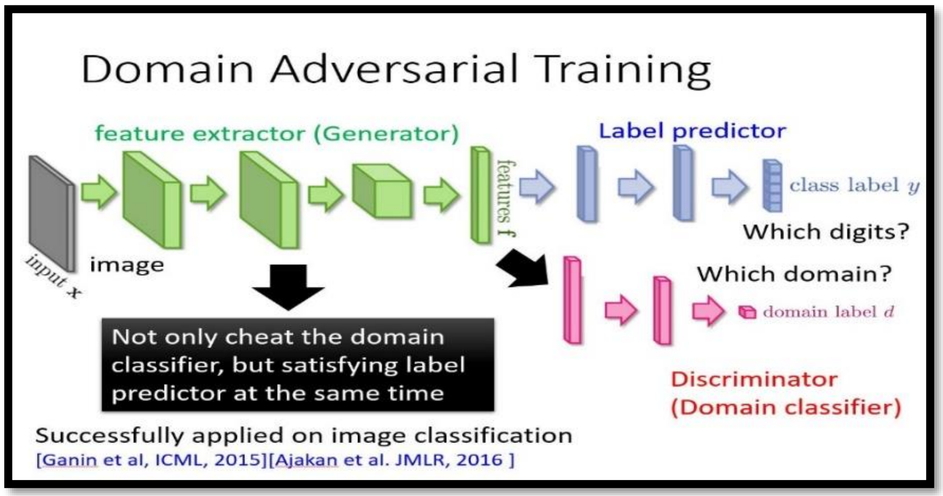
Speech Signal Generation (Regression Task)



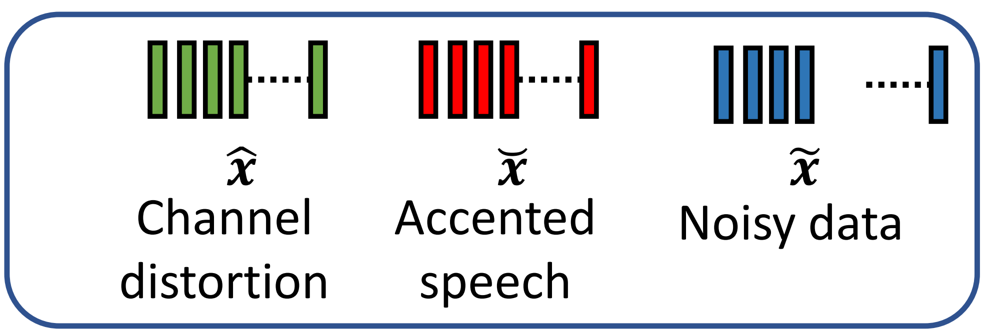


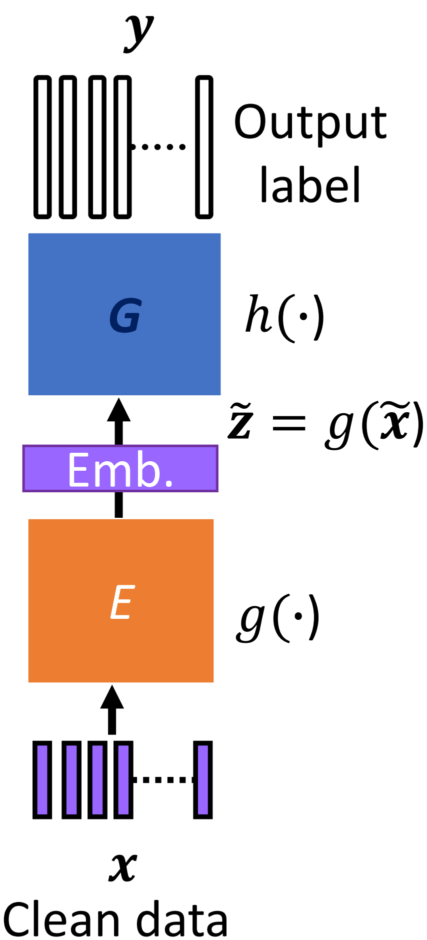


Speech, Speaker, Emotion Recognition and Lip-reading (Classification Task)



Acoustic Mismatch





More GANs in Speech

Diagnosis of autism spectrum
Jun Deng, Nicholas Cummins, Maximilian Schmitt, Kun Qian, Fabien Ringeval, and Björn Schuller, Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations, ACM DH, 2017.

Emotion recognition
Jonathan Chang, and Stefan Scherer, Learning Representations of Emotional Speech with Deep Convolutional Generative Adversarial Networks, ICASSP, 2017.

Robust ASR
Dmitriy Serdyuk, Kartik Audhkhasi, Philémon Brakel, Bhuvana Ramabhadran, Samuel Thomas, and Yoshua Bengio, Invariant Representations for Noisy Speech Recognition, arXiv, 2016.

Speaker verification
Hong Yu, Zheng-Hua Tan, Zhanyu Ma, and Jun Guo, Adversarial Network Bottleneck Features for Noise Robust Speaker Verification, arXiv, 2017.

References

Speech enhancement (conventional methods)

- Yuxuan Wang and Deliang Wang, Cocktail Party Processing via Structured Prediction, NIPS 2012.
- Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, An Experimental Study on Speech Enhancement Based on Deep Neural Networks," IEEE SPL, 2014.
- Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, A Regression Approach to Speech Enhancement Based on Deep Neural Networks, IEEE/ACM TASLP, 2015.
- Xugang Lu, Yu Tsao, Shigeki Matsuda, Chiori Hori, Speech Enhancement Based on Deep Denoising Autoencoder, Interspeech 2012.
- Zhuo Chen, Shinji Watanabe, Hakan Erdogan, John R. Hershey, Integration of Speech Enhancement and Recognition Using Long-short term Memory Recurrent Neural Network, Interspeech 2015.
- Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R. Hershey, and Bjorn Schuller, Speech Enhancement with LSTM Recurrent Neural Networks and Its Application to Noise-robust ASR, LVA/ICA, 2015.
- Szu-Wei Fu, Yu Tsao, and Xugang Lu, SNR-aware Convolutional Neural Network Modeling for Speech Enhancement, Interspeech, 2016.
- Szu-Wei Fu, Yu Tsao, Xugang Lu, and Hisashi Kawai, End-to-end Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks, arXiv, IEEE/ACM TASLP, 2018.

Speech enhancement (GAN-based methods)

- Pascual Santiago, Bonafonte Antonio, and Serra Joan, SEGAN: Speech Enhancement Generative Adversarial Network, Interspeech, 2017.
- Michelsanti Daniel, and Zheng-Hua Tan, Conditional Generative Adversarial Networks for Speech Enhancement and Noise-robust Speaker Verification, Interspeech, 2017.
- Donahue Chris, Li Bo, and Prabhavalkar Rohit, Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition, ICASPP, 2018.
- Higuchi Takuya, Kinoshita Keisuke, Delcroix Marc, and Nakatani Tomohiro, Adversarial Training for Data-driven Speech Enhancement without Parallel Corpus, ASRU, 2017.

References

Postfilter (conventional methods)

- Toda Tomoki, and Tokuda Keiichi, A Speech Parameter Generation Algorithm Considering Global Variance for HMM-based Speech Synthesis, IEICE Trans. Inf. Syst., 2007.
- Sil'en Hanna, Helander Elina, Nurminen Jani, and Gabbouj Moncef, Ways to Implement Global Variance in Statistical Speech Synthesis, Interspeech, 2012.
- Takamichi Shinnosuke, Toda Tomoki, Neubig Graham, Sakti Sakriani, and Nakamura Satoshi, A Postfilter to Modify the Modulation Spectrum in HMM-based Speech Synthesis, ICASSP, 2014.
- Ling-Hui Chen, Tuomo Raitio, Cassia Valentini-Botinhao, Junichi Yamagishi, and Zhen-Hua Ling, DNN-based Stochastic Postfilter for HMM-based Speech Synthesis, Interspeech, 2014.
- Ling-Hui Chen, Tuomo Raitio, Cassia Valentini-Botinhao, Zhen-Hua Ling, and Junichi Yamagishi, A Deep Generative Architecture for Postfiltering in Statistical Parametric Speech Synthesis, IEEE/ACM TASLP, 2015.

Postfilter (GAN-based methods)

- Kaneko Takuhiro, Kameoka Hirokazu, Hojo Nobukatsu, Ijima Yusuke, Hiramatsu Kaoru, and Kashino Kunio, Generative Adversarial Network-based Postfilter for Statistical Parametric Speech Synthesis, ICASSP, 2017.
- Kaneko Takuhiro, Takaki Shinji, Kameoka Hirokazu, and Yamagishi Junichi, Generative Adversarial Network-Based Postfilter for STFT Spectrograms, Interspeech, 2017.
- Saito Yuki, Takamichi Shinnosuke, and Saruwatari Hiroshi, Training Algorithm to Deceive Anti-spoofing Verification for DNN-based Speech Synthesis, ICASSP, 2017.
- Saito Yuki, Takamichi Shinnosuke, Saruwatari Hiroshi, Saito Yuki, Takamichi Shinnosuke, and Saruwatari Hiroshi, Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks, IEEE/ACM TASLP, 2018.
- Bajibabu Bollepalli, Lauri Juvela, and Alku Paavo, Generative Adversarial Network-based Glottal Waveform Model for Statistical Parametric Speech Synthesis, Interspeech, 2017.
- Yang Shan, Xie Lei, Chen Xiao, Lou Xiaoyan, Zhu Xuan, Huang Dongyan, and Li Haizhou, Statistical Parametric Speech Synthesis Using Generative Adversarial Networks Under a Multi-task Learning Framework, ASRU, 2017.

References

VC (conventional methods)

- Toda Tomoki, Black Alan W, and Tokuda Keiichi, Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory, IEEE/ACM TASLP, 2007.
- Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai, Voice Conversion Using Deep Neural Networks with Layer-wise Generative Training, IEEE/ACM TASLP, 2014.
- Srinivas Desai, Alan W Black, B. Yegnanarayana, and Kishore Prahallad, Spectral mapping Using artificial Neural Networks for Voice Conversion, IEEE/ACM TASLP, 2010.
- Nakashika Toru, Takiguchi Tetsuya, Ariki Yasuo, High-order Sequence Modeling Using Speaker-dependent Recurrent Temporal Restricted Boltzmann Machines for Voice Conversion, Interspeech, 2014.
- Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino, Sequence-to-sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks, Interspeech, 2017.
- Zhizheng Wu, Tuomas Virtanen, Eng-Siong Chng, and Haizhou Li, Exemplar-based Sparse Representation with Residual Compensation for Voice Conversion, IEEE/ACM TASLP, 2014.
- Szu-Wei Fu, Pei-Chun Li, Ying-Hui Lai, Cheng-Chien Yang, Li-Chun Hsieh, and Yu Tsao, Joint Dictionary Learning-based Non-negative Matrix Factorization for Voice Conversion to Improve Speech Intelligibility After Oral Surgery, IEEE TBME, 2017.
- Yi-Chiao Wu, Hsin-Te Hwang, Chin-Cheng Hsu, Yu Tsao, and Hsin-Min Wang, Locally Linear Embedding for Exemplar-based Spectral Conversion, Interspeech, 2016.

VC (GAN-based methods)

- Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks, arXiv, 2017.
- Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino, Sequence-to-sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks, Interspeech, 2017.
- Takuhiro Kaneko, and Hirokazu Kameoka. Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks, arXiv, 2017.

References

ASR

- Yusuke Shinohara, Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition. Interspeech, 2016.
- Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie, Domain Adversarial Training for Accented Speech Recognition, ICASSP, 2018
- Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, Cross-domain Speech Recognition Using Nonparallel Corpora with Cycle-consistent Adversarial Networks, ASRU, 2017.
- Anuroop Sriram, Heewoo Jun, Yashesh Gaur, and Sanjeev Satheesh, Robust Speech Recognition Using Generative Adversarial Networks, arXiv, 2017.

Speaker recognition

- Qing Wang, Wei Rao, Sining Sun, Lei Xie, Eng Siong Chng, and Haizhou Li, Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition, ICASSP, 2018.

Emotion recognition

- Saurabh Sahu, Rahul Gupta, Ganesh Sivaraman, Wael AbdAlmageed, and Carol Espy-Wilson, Adversarial Auto-encoders for Speech Based Emotion Recognition. Interspeech, 2017.

Lipreading

- Michael Wand, and Jürgen Schmidhuber, Improving Speaker-Independent Lipreading with Domain-Adversarial Training, arXiv, 2017.

GANs in ICASSP 2018

- Saurabh Sahu, Rahul Gupta, Ganesh Sivaraman, Carol Espy-Wilson, Smoothing Model Predictions using Adversarial Training Procedures for Speech Based Emotion Recognition
- Fuming Fang, Junichi Yamagishi, Isao Echizen, Jaime Lorenzo-Trueba, High-quality Nonparallel Voice Conversion Based on Cycle-consistent Adversarial Network
- Lauri Juvela, Bajibabu Bollepalli, Xin Wang, Hirokazu Kameoka, Manu Airaksinen, Junichi Yamagishi, Paavo Alku, Speech Waveform Synthesis from MFCC Sequences with Generative Adversarial Networks
- Zhong Meng, Jinyu Li, Yifan Gong, Biing-Hwang (Fred) Juang, Adversarial Teacher-Student Learning for Unsupervised Domain Adaptation
- Zhong Meng, Jinyu Li, Zhuo Chen, Yong Zhao, Vadim Mazalov, Yifan Gong, Biing-Hwang (Fred) Juang, Speaker-Invariant Training via Adversarial Learning
- Sen Li, Stephane Villette, Pravin Ramadas, Daniel Sinder, Speech Bandwidth Extension using Generative Adversarial Networks
- Qing Wang, Wei Rao, Sining Sun, Lei Xie, Eng Siong Chng, Haizhou Li, Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition
- Hu Hu, Tian Tan, Yanmin Qian, Generative Adversarial Networks Based Data Augmentation for Noise Robust Speech Recognition
- Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari, Text-to-speech Synthesis using STFT Spectra Based on Low-/multi-resolution Generative Adversarial Networks
- Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Ye Bai, Adversarial Multilingual Training for Low-resource Speech Recognition
- Meet H. Soni, Neil Shah, Hemant A. Patil, Time-frequency Masking-based Speech Enhancement using Generative Adversarial Network
- Taira Tsuchiya, Naohiro Tawara, Tetsuji Ogawa, Tetsunori Kobayashi, Speaker Invariant Feature Extraction for Zero-resource Languages with Adversarial Learning

GANs in ICASSP 2018

- Jing Han, Zixing Zhang, Zhao Ren, Fabien Ringeval, Bjoern Schuller, Towards Conditional Adversarial Training for Predicting Emotions from Speech
- Chenxing Li, Lei Zhu, Shuang Xu, Peng Gao, Bo Xu, CBLDNN-based Speaker-independent Speech Separation via Generative Adversarial Training
- Anuroop Sriram, Heewoo Jun, Yashesh Gaur, Sanjeev Satheesh, Robust Speech Recognition using Generative Adversarial Networks
- Cem Subakan, Paris Smaragdis, Generative Adversarial Source Separation,
- Ashutosh Pandey, Deliang Wang, On Adversarial Training and Loss Functions for Speech Enhancement
- Bin Liu, Shuai Nie, Yaping Zhang, Dengfeng Ke, Shan Liang, Wenju Liu, Boosting Noise Robustness of Acoustic Model via Deep Adversarial Training
- Yang Gao, Rita Singh, Bhiksha Raj, Voice Impersonation using Generative Adversarial Networks
- Aditay Tripathi, Aanchan Mohan, Saket Anand, Maneesh Singh, Adversarial Learning of Raw Speech Features for Domain Invariant Speech Recognition
- Zhe-Cheng Fan, Yen-Lin Lai, Jyh-Shing Jang, SVSGAN: Singing Voice Separation via Generative Adversarial Network
- Santiago Pascual, Maruchan Park, Joan Serra, Antonio Bonafonte, Kang-Hun Ahn, Language and Noise Transfer in Speech Enhancement Generative Adversarial Network

A promising research direction and still has room for further improvements in the speech signal processing domain

Thank You Very Much

Tsao, Yu   Ph.D.
yu.tsao@citi.sinica.edu.tw
https://www.citi.sinica.edu.tw/pages/yu.tsao/contact_zh.html