Towards noise-adaptive training of GANs: adding noise information to the objective function

Justina Ramonaitė, Gražina Korvel
Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania

Abstract

Speech enhancement is a broad research field that covers the improvement of speech signals by reducing or completely removing various distortions and improving the intelligibility and quality of the signal. Different approaches have been proposed, varying from traditional methods based on filtering or other mathematical and statistical operations to deep learning. The latter includes generative adversarial networks (GANs). The first GAN-type models used for speech enhancement were introduced in 2017 and have improved greatly since then. They have demonstrated significant improvement in the most common quality evaluation metrics, such as PESQ and STOI.

Goal

This study investigates the possibility of improving the performance of GANs by including noise-related data into the training objective. **The main focus of this research is on developing a loss function that would adjust according to the selected properties of initial noise.**

Models

- **SEGAN:** one of the 1st GANs for speech enhancement, introduced in 2017.
- **CMGAN:** conformer-based metric GAN, introduced in 2022.

SI-SNR and adjusted loss

Scale-invariant signal-to-noise ratio (SI-SNR) loss is defined as

$$L_{SI-SNR} = -10\log_{10}\left(\frac{\|s_{target}\|^2}{\|e_{noise}\|^2}\right),$$

where $s_{target} = \frac{(\hat{s}^T s)}{\|s\|^2} s$, $e_{noise} = \hat{s} - s_{target}$, s and $\hat{s}$ are clean and enhanced speech signals respectively.

Then the new loss is defined as

$$L_{new} = L_{old} + \gamma L_{SI-SNR},$$

where L_{old} is the original loss that differs for each model.

This work explores weight γ, that is constructed of 3 components

$$\gamma(x) = \alpha \begin{cases} f(x), & \text{if noise type } \in C; \\ cf(x), & \text{otherwise.} \end{cases}$$

Here α controls the overall size of the weight, so it aligns with the rest of the loss, $f(x)$ is a function that depends on initial signal-to-noise ratio, c is selected coefficient and C represents the selected noise types. In this work $C = \{\text{"car"}\}$ and $cf(x)$ does not exceed 1.

Spectral characteristics

Set	Noise	Centroid	Bandwidth	Flatness	Rolloff
Test	DLIVING	7242	7846	0.02754	18083
	OOFFICE	5379	7373	0.00478	15544
	SCAFE	3508	5138	0.00758	8536
	SPSQUARE	3121	5380	0.00327	7510
	TBUS	2221	4257	0.00114	4928
Train	DKITCHEN	8306	5706	0.04549	14177
	OMEETING	7376	7670	0.04021	17714
	PCAFETER	4159	5678	0.01182	10072
	PRESTO	4486	5086	0.02313	9820
	PSTATION	2542	4180	0.00317	5562
	STRAFFIC	3398	4935	0.00497	8497
	TCAR	833	3275	0.00004	348
	TMETRO	2358	4031	0.00200	5257

Results

Model	c	PESQ	STOI	CSIG	CBAK	COVL
Noisy test set		*1.97*	*0.92*	*3.35*	*2.44*	*2.63*
SEGAN	-	2.19	0.93	3.53	3.01	2.84
	2	2.36	0.93	3.73	3.08	3.03
	4	2.44	0.93	3.76	3.09	3.08
CMGAN	-	2.97	0.95	4.23	3.08	3.61
	2	3.24	0.96	4.47	3.27	3.88
	4	3.25	0.96	4.48	3.26	3.89

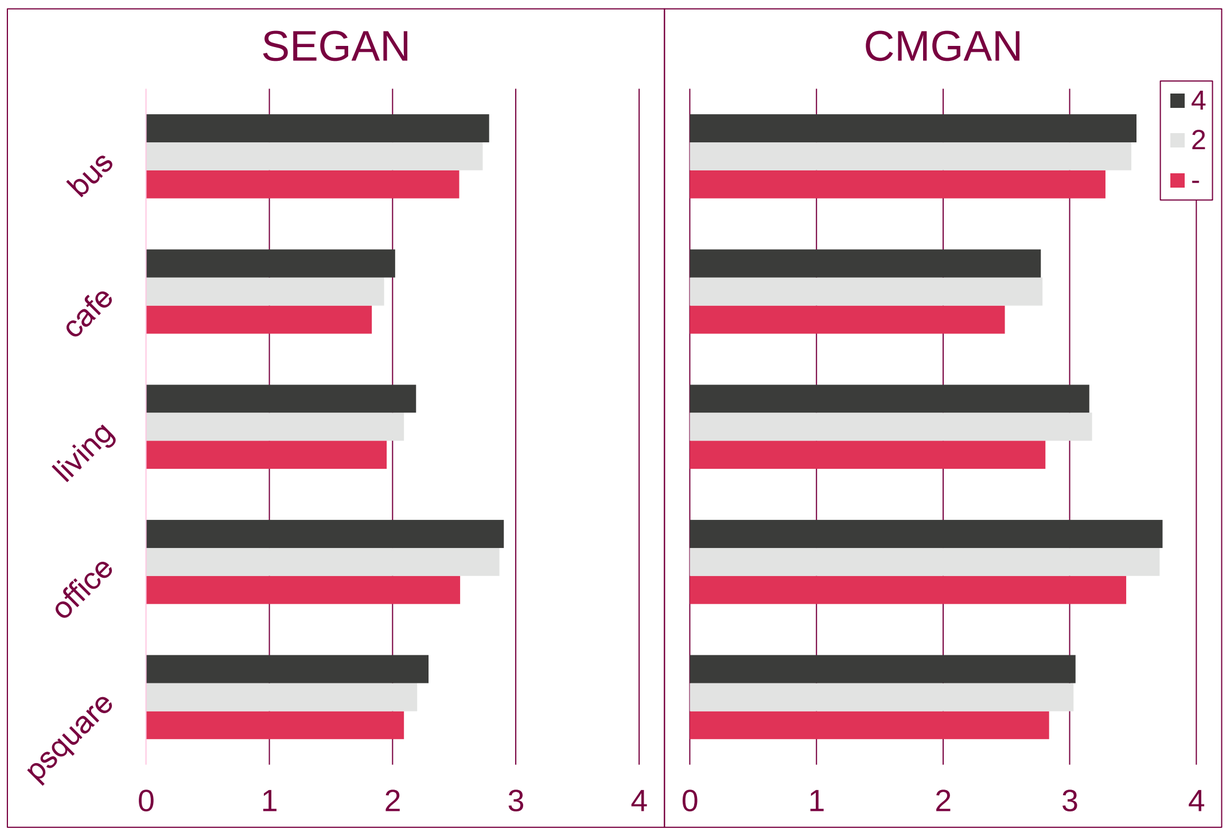

Fig. 1: PESQ with different c values per noise type.

Conclusions and future works

The experiments show promising results, maintaining the improvement of the models and also revealing possible further directions toward more elaborate noise evaluation and profiling.

Acknowledgements

This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-24-118.

The authors are also thankful for the high-performance computing resources provided by the Information Technology Research Center of Vilnius University.