BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Speech Enhancement using Neural Audio Codecs

Dominik Klement*, Supervisor: doc. Lukas Burget, Ph.D.

Abstract
Speech enhancement aims to remove background noise from speech signals while preserving speech quality and intelligibility. In this work, we propose a novel dual-branch model based on neural audio codec that separates clean speech and noise into two separate streams. In order to allow unsupervised training, we combine the branches and force the output to resemble the input noisy speech. Our experiments show that supervised models outperform strong baselines in SI-SDR and achieve competitive perceptual scores, while our unsupervised model significantly improve noisy inputs without requiring paired data. These results demonstrate the potential of our approach for both supervised and unsupervised speech enhancement, contributing towards more generalizable and robust systems.

*xkleme15@vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Speech Enhancement (SE) aims to remove background noise from speech while preserving the underlying clean signal. Historically, SE methods were based on classical signal processing techniques. Over the past decade, however, deep learning-based end-to-end (E2E) models have become the dominant approach.

Training E2E SE models with real-world data is challenging due to the absence of clean-noisy speech pairs. Consequently, models are typically trained on simulated mixtures created by adding noise to clean speech [1, 2]. Although effective, these mixtures poorly reflect real-world complexities, limiting model generalization.

To address this, unsupervised SE methods [3, 4, 5] have been proposed, aiming to enhance speech without requiring clean references. These methods assume uncorrelated noise and speech or attempt to model clean speech distributions directly. However, maintaining consistency between the noisy input and the enhanced output (i.e. preserving uttered content, speaker identity, prosody, to name a few) remains a significant challenge.

In this work, we propose a novel dual-branch architecture based on neural audio codecs (NACs), which separates clean speech and noise into two distinct audio streams. By reconstructing the original noisy input from the sum of the two outputs, we ensure consistency and enable unsupervised training by utilizing clean speech and noise discriminators to guide the two branches.

2. Proposed Method

The key idea behind our approach is as follows: If we have a model with two separate audio output streams and we encourage one to resemble clean speech, then summing the two outputs and forcing the result to closely match the original noisy input should enforce consistency between the noisy input and the enhanced speech.

Let $\mathbf{x} \in \mathbb{R}^T$ be the input noisy speech signal of length T. As depicted in Figure 3, we define a convolutional encoder $E : \mathbb{R}^T \to \mathbb{R}^{N \times d}$ and a decoder $D : \mathbb{R}^{N \times d} \to \mathbb{R}^T$, where N is the number of frames, inspired by the Descript Audio Codec (DAC) [6]. These networks map raw waveforms into high-dimensional latent representations and reconstruct them back.

Passing $\mathbf{x}$ through E yields a latent sequence $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^{N \times d}$, which captures local structure. To allow the model to capture longer-term dependencies like prosody and speaker characteristics, as well as structured noise patterns (e.g., sirens, engines), we use two separate transformers with rotary positional embeddings (roformer) [7]: R_{CS} and R_N.

These transformers produce two separate latent sequences, $\mathbf{z}_{CS} = R_{CS}(\mathbf{z})$ and $\mathbf{z}_N = R_N(\mathbf{z})$. Each latent stream is then quantized using residual vector quantization (RVQ), producing quantized representations $\hat{\mathbf{z}}_{CS}$ and $\hat{\mathbf{z}}_N$. The quantization allows us to control the information bandwidth of each branch and reduces information leakage between them.

Both branches are then decoded separately to produce waveform outputs $\hat{\mathbf{x}}_{CS} = D(\hat{\mathbf{z}}_{CS})$ and $\hat{\mathbf{x}}_N = D(\hat{\mathbf{z}}_N)$, corresponding to the estimated clean speech and noise, respectively.

To account for possible amplitude mismatches, we compute optimal scalars α and β by solving:

$$\alpha^*, \beta^* = \arg\min_{\alpha,\beta} \|x - \alpha\hat{\mathbf{x}}_{CS} - \beta\hat{\mathbf{x}}_N\|_2^2. \quad (1)$$

The reconstructed noisy input is then obtained as:

$$\hat{\mathbf{x}} = \alpha\hat{\mathbf{x}}_{CS} + \beta\hat{\mathbf{x}}_N. \quad (2)$$

To optimize the model, we treat the system as 3 separate generative adversarial networks (GANs) [8]; hence, employing 3 discriminators: D_{CS}, D_N, D_{NS} for clean speech, noise, and noisy speech respectively. The architecture of each discriminator is derived from NACs—i.e., an ensemble of 8 discriminators operating on a complex Short Time Fourier Transform (STFT) spectrogram with different window sizes and hop lengths, originally employed in NACs to increase the fidelity of the reconstructed audio. The discriminator is trained against generator (in our case a model producing audio) to produce a score close to 1 if the discriminator input is a real sample, or close to 0 if it is produced by the generator (i.e. fake). It has been proved by [8] that after convergence, the generator will become a sampler from the real data distribution.

These discriminators ensure that each branch learns the correct distribution: clean speech for D_{CS}, noise for D_N, and high-quality reconstruction for D_{NS}.

Finally, to stabilize training and further enhance quality, we add a reconstruction loss combining SI-SDR and mel-spectrogram distance:

$$\mathcal{L}_r = \text{SI-SDR}(\mathbf{x}, \hat{\mathbf{x}}) + \|\text{logmel}(\mathbf{x}) - \text{logmel}(\hat{\mathbf{x}})\|_1. \quad (3)$$

3. Experiment Setup

We trained our models on a dataset combining speech, noise, and room impulse responses (RIRs), following the URGENT challenge setup [9]. The training corpus includes 2500 hours of speech, 500 hours of various noise types, and more than 60,000 RIRs. All audio data is resampled to 16kHz.

We train the models in 3 steps: pre-train E, R_{CS}, D, and D_{CS} to perform SE using simulated mixtures using $\mathcal{L}_r$ between the ground-truth and the estimated clean speech, step 2: introduce R_N, D_N, and train the entire 2-branch model to perform SE and noisy speech reconstruction, step 3: transition to fully unsupervised training by removing clean-speech $\mathcal{L}_r$, relying only on adversarial objectives and reconstruction consistency.

We validate the models on a test subset of a well-established noisy speech dataset VCTK-Demand [10], using both, signal-based metric SI-SDR [11], and perceptual metrics PESQ [12], STOI [13], DNS-MOS [14], and UTMOS [15].

4. Experiments

Table 1 shows the comparison of strong baselines MetricGan+ [16], HiFi-GAN-2 [17], and FINALLY [2] with our models. It can be seen that our supervised models perform the best in SI-SDR, and achieve the second best STOI scores. Furthermore, our models are competitive in the other 3 perceptual metrics, namely DNS-MOS and UTMOS, achieving the second best results.

Although the 2-branch unsupervised model lacks behind the supervised models, it still outperforms the noisy input lowerbounds. The decrease of performance is attributed to the slight leakage of noise to the clean speech branch, as the clean-speech discriminator D_{CS} does not enforce fine-grained details preservation strongly, allowing the noisy speech reconstruction gradients to overrule gradients coming from D_{CS}. However, without D_{CS}, the clean speech branch leaks the entire noise, proving its necessity.

Additionally, the noise branch accurately models residual noise, validated by strong noisy input reconstruction scores showed in Table 2, and a sample depicted in Figure 4. We observed that the noise discriminator D_N plays crucial role in preventing clean speech leakage into the noise branch, which in turn results in better quality of clean speech.

5. Conclusion

In this work, we introduced a novel dual-branch neural audio codec-based model for speech enhancement. By reconstructing the input noisy speech, our method enforces consistency and enables both, supervised and unsupervised training.

Our supervised models outperform strong baselines in SI-SDR and achieve competitive scores across several perceptual metrics. Although the unsupervised variant performs slightly worse, it still demonstrates a clear enhancement over the noisy inputs, validating the approach.

References

[1] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative adversarial network, 2017.

[2] Nicholas Babaev, Kirill Tamogashev, Azat Saginbaev, Ivan Shchekotov, Hanbin Bae, Hosang Sung, WonJun Lee, Hoon-Young Cho, and Pavel Andreev. Finally: fast and universal speech enhancement with studio-like quality, 2024.

[3] N. Alamdari, A. Azarang, and N. Kehtarnavaz. Improving deep speech denoising by noisy2noisy signal mapping. *Applied Acoustics*, 172:107631, 2021.

[4] Madhav Mahesh Kashyap, Anuj Tambwekar, Krishnamoorthy Manohara, and S. Natarajan. Speech denoising without clean training data: A noise2noise approach. In *Interspeech 2021*, interspeech_2021. ISCA, 2021.

[5] Szu-Wei Fu, Cheng Yu, Kuo-Hsuan Hung, Mirco Ravanelli, and Yu Tsao. Metricgan-u: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7412–7416, 2022.

[6] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan, 2023.

[7] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[9] Wangyou Zhang, Robin Scheibler, Kohei Saijo, Samuele Cornell, Chenda Li, Zhaoheng Ni, Jan Pirklbauer, Marvin Sach, Shinji Watanabe, Tim Fingscheidt, and Yanmin Qian. Urgent challenge: Universality, robustness, and generalizability for speech enhancement. In *Interspeech 2024*, interspeech_2024, page 4868–4872. ISCA, September 2024.

[10] Cassia Valentini-Botinhao. Noisy speech database for training speech enhancement algorithms and TTS models, 2017.

[11] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630, 2019.

[12] 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 *Proceedings of the Acoustics, Speech, and Signal Processing, 200. on IEEE International Conference - Volume 02*, ICASSP '01, page 749–752, USA, 2001. IEEE Computer Society.

[13] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

[14] Chandan K A Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, 2021.

[15] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022, 2022.

[16] Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. Metricgan+: An improved version of metricgan for speech enhancement, 2021.

[17] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features. In *WASPAA 2021*, October 2021.