SPEECH ENHANCEMENT USING NEURAL AUDIO CODEC

Dominik Klement supervised by doc. Lukáš Burget, Ph.D.

BRNO FACULTY OF INFORMATION TECHNOLOGY
UNIVERSITY OF TECHNOLOGY

WHAT IS SPEECH ENHANCEMENT?

An algorithm suppresses sounds other than clean speech.
Currently, a machine learning model.

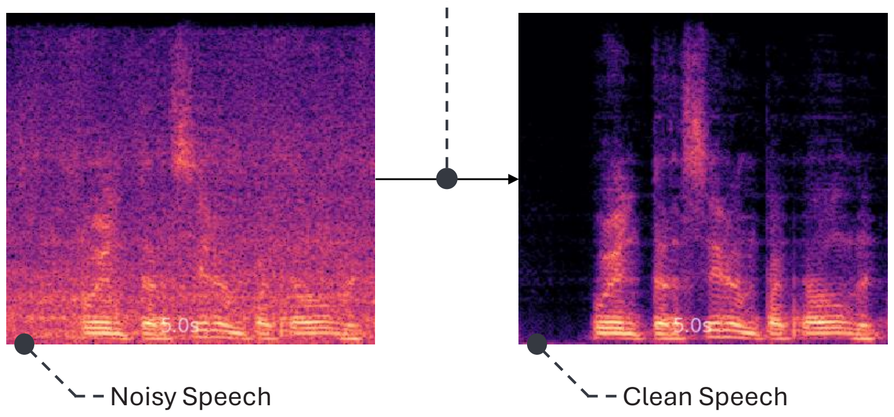
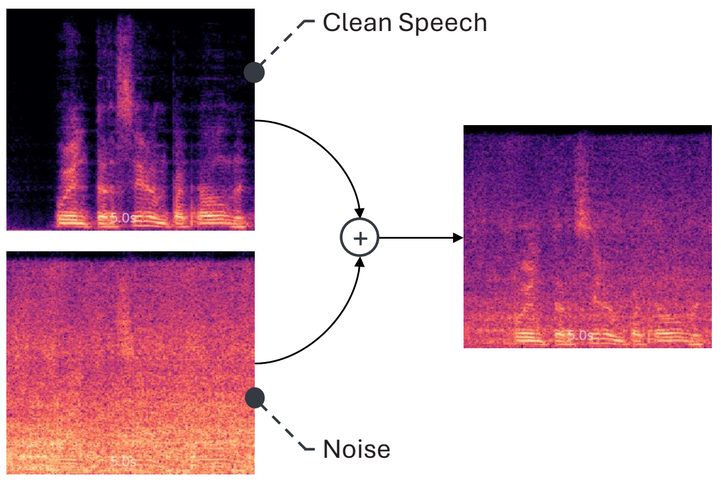


Figure 1: Spectrogram of noisy speech and its clean speech counterpart.

HOW IS IT USUALLY TRAINED?



- Noisy speech is created by summing a clean speech and a noise waveforms.
- A model is trained to map noisy speech to the original clean speech.
- This creates a discrepancy between training and inference, creating a possible generalization gap.

Figure 2: Example of simulated mixture creation process.

PROPOSED SOLUTION

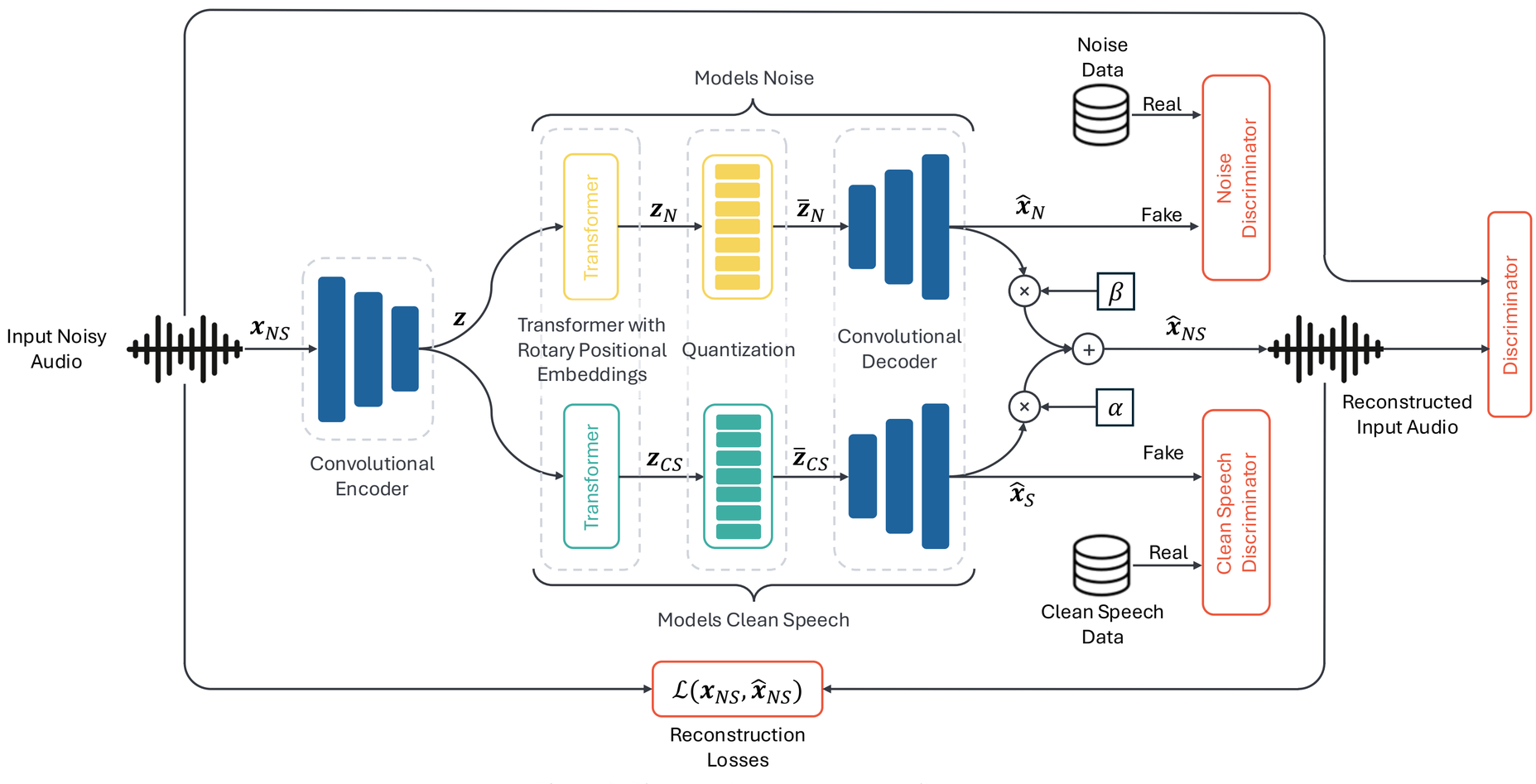


Figure 3: Diagram of the proposed architecture.

RESULTS

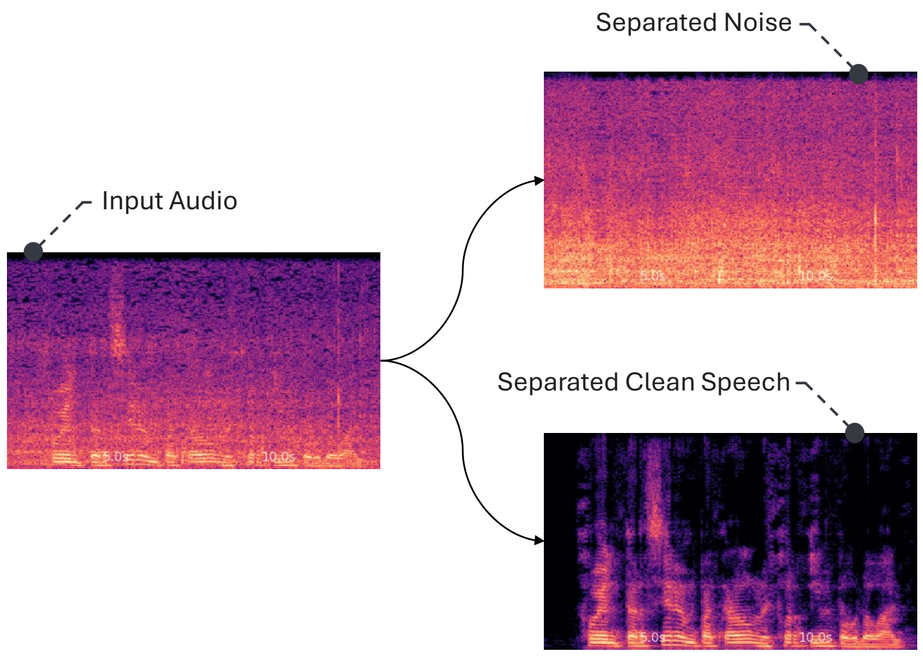


Figure 4: Example of separated audio by 2-branch model.

Table 1: Comparison of baselines and our approaches. All but the last row are trained in a supervised manner.

Method	DNS-MOS↑	UTMOS↑	PESQ↑	STOI↑	SI-SDR↑
Noisy Input	2.53	2.62	1.98	0.92	8.4
Ground Truth	3.22	4.07	-	-	-
MetricGAN+	2.95	3.62	**3.14**	0.93	8.6
HiFi-GAN-2	3.12	3.99	**3.14**	**0.95**	17.9
FINALLY	**3.22**	**4.32**	2.94	0.92	4.6
1-branch* (ours)	3.15	4.00	2.86	0.94	**19.3**
2-branch* (ours)	3.17	3.91	2.82	0.94	**19.3**
2-branch unsupervised	3.02	3.60	2.25	0.92	15.06

Table 2: Noisy speech reconstruction results for supervised and unsupervised 2-branch models.

Method	PESQ↑	STOI↑	SI-SDR↑
Supervised	4.17	0.98	22.7
Unsupervised	4.22	0.98	23.91