Unsupervised speech enhancement with deep dynamical generative speech and noise models

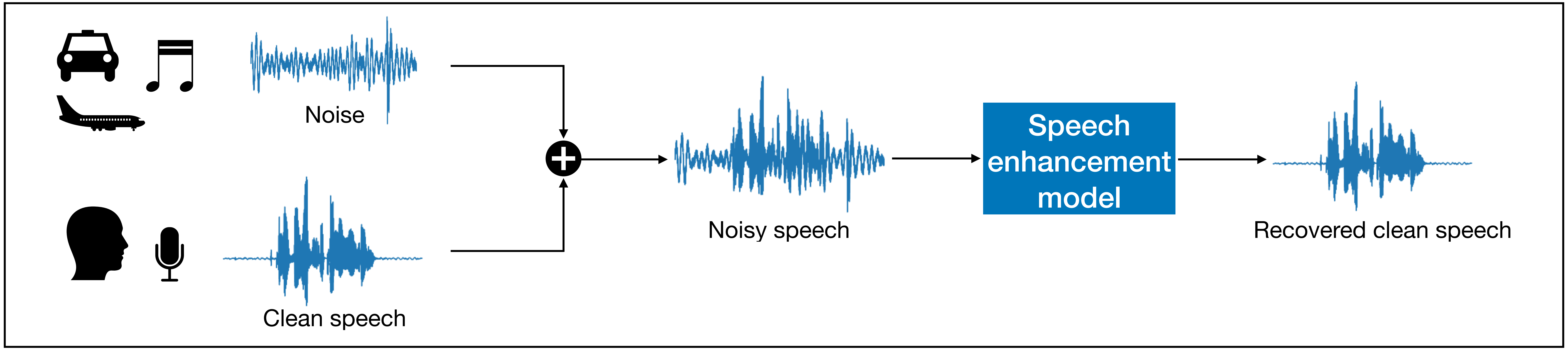
Xiaoyu Lin[1], Simon Leglaive[2], Laurent Girin[3], Xavier Alameda-Pineda[1]

[1] Inria Grenoble Rhône-Alpes, Univ. Grenoble Alpes, France
[2] CentraleSupélec, IETR (UMR CNRS 6164), France
[3] Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab, France

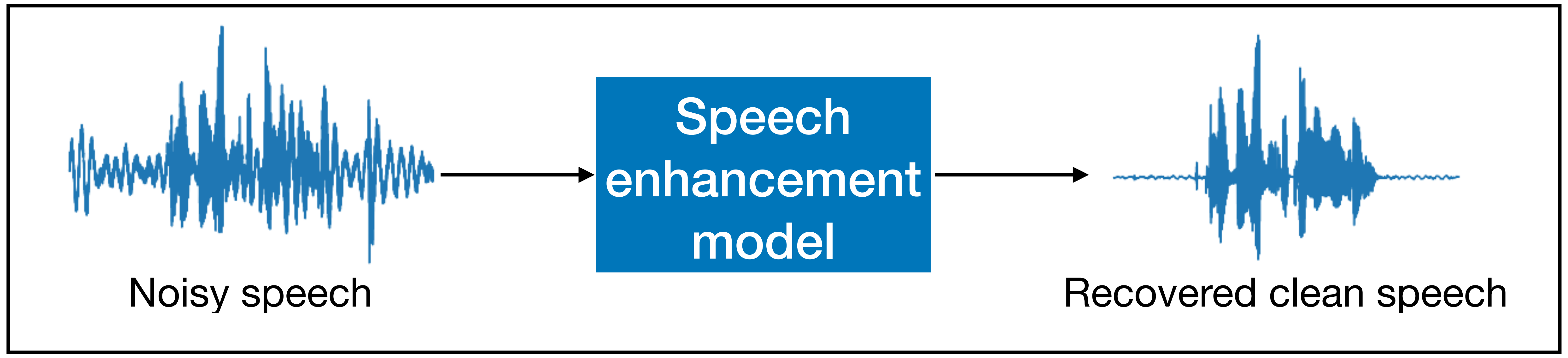
INTERSPEECH 2023

Context

Speech enhancement



Supervised speech enhancement methods



- Direct mapping from the noisy speech to the recovered clean speech.
- Model trained by minimizing a certain distance between the ground truth and estimated clean speech.

Context

Limitations of the supervised speech enhancement methods

- Requirements for large amount of parallel clean-noisy speech signals for training.
- Poor generalization ability to noise types and acoustic conditions that were not seen during training.

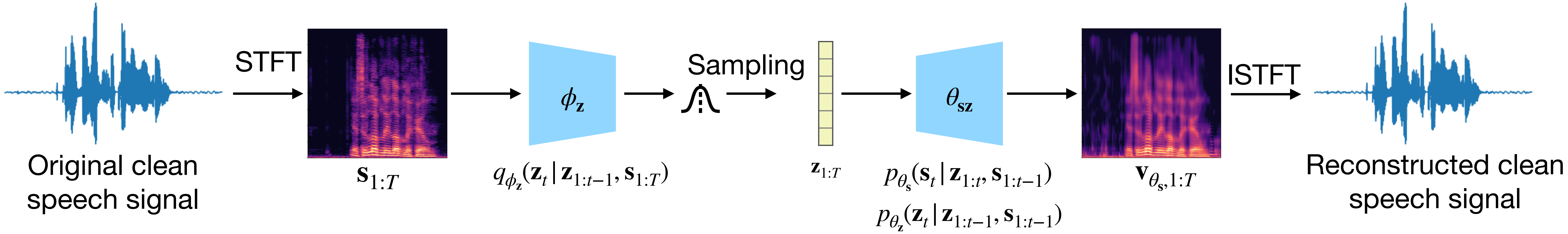
Unsupervised speech enhancement methods

- No need for parallel clean-noisy speech dataset for training.
- Can be further divided into **unsupervised noise-dependent (U-ND)** and **unsupervised noise-agnostic (U-NA)** methods.
- **Unsupervised noise-dependent (U-ND)** methods use noise or noisy samples during training.
- **Unsupervised noise-agnostic (U-NA)** methods estimate the noise characteristics directly at test time.

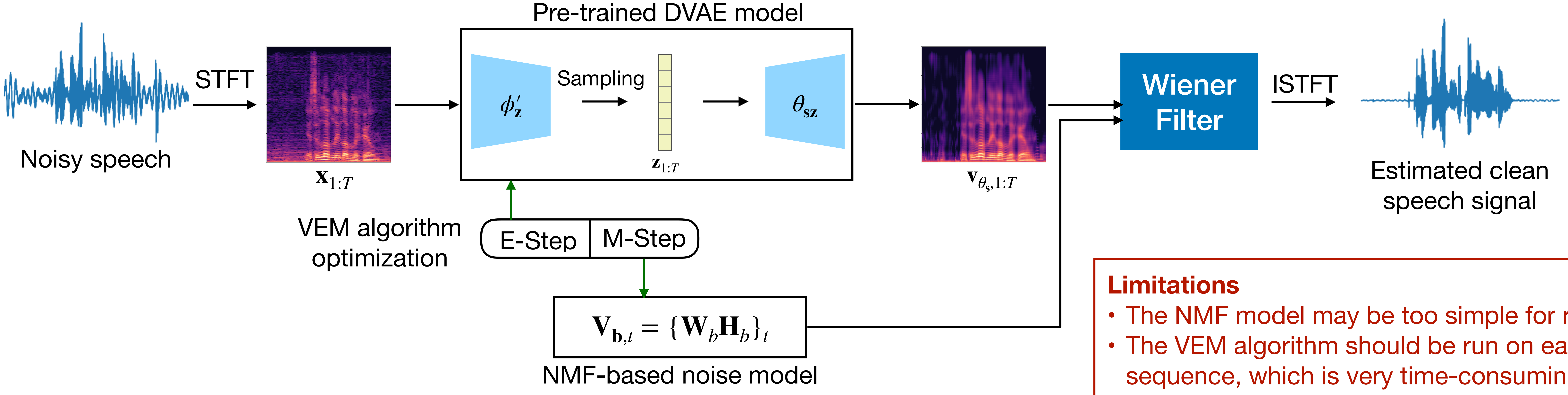
Context

Deep probabilistic model-based U-NA method: RVAE-VEM model

- Pre-training on clean speech signals with Dynamical VAE (DVAE) model



- Speech enhancement with the pre-trained DVAE model and NMF-based noise model



Limitations

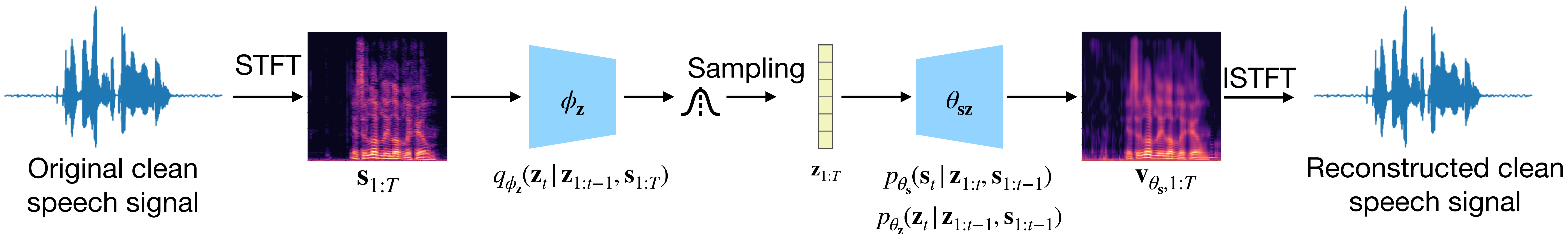
- The NMF model may be too simple for real-world noise.
- The VEM algorithm should be run on each test noisy sequence, which is very time-consuming.

Main contributions

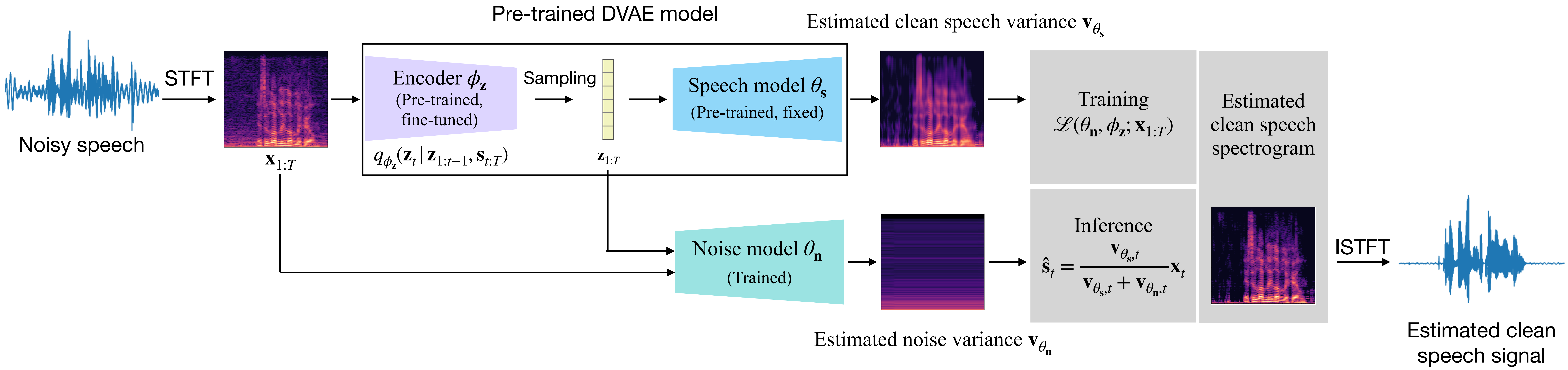
- Replace the NMF noise model with a deep dynamical generative model (DDGM).
  - The DDGM model is a general class of dynamical models for the generation of sequential data based on DNNs.
- Implement and test the DDGM noise model with different variable dependencies.
  - We test the noise model with three kind of dependencies: the DVAE latent variables (LV), or the noisy observations (NO), or both (NOLV).
- Flexible to be trained in different configurations.
  - The proposed method can be trained in both the U-NA and the U-ND configurations. Further, it can be first trained in U-ND configuration, then fine-tuned in U-NA configuration.
- Performance comparable to that of the NMF-based method with less inference time in U-ND configuration.
  - The proposed method requires much less computation time during inference when trained and tested in the U-ND configuration.

DDGM-based speech enhancement method

- Pre-training on clean speech signals with Dynamical VAE (DVAE) model



- Speech enhancement with the pre-trained DVAE model and DDGM-based noise model



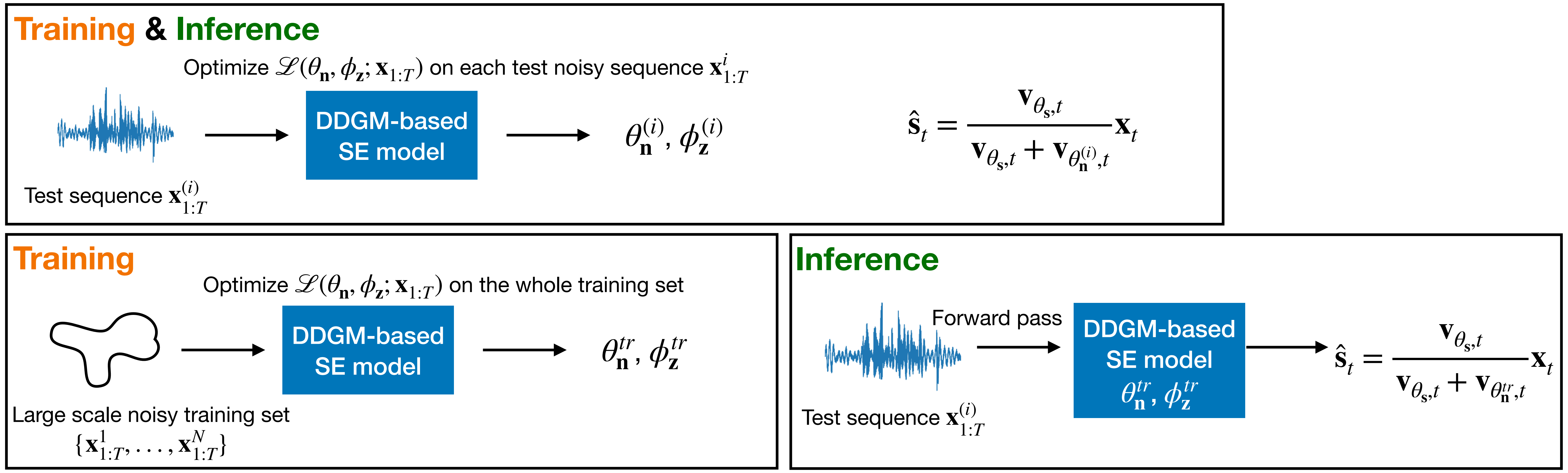
Different variable dependencies and training configurations

Three variable dependencies of the noise model

- **DVAE latent variables (LV):** $\mathbf{v}_{\theta_\mathbf{n},t} = \mathbf{v}_{\theta_\mathbf{n},t}(\mathbf{z}_{1:T})$
- **Noisy observations (NO):** $\mathbf{v}_{\theta_\mathbf{n},t} = \mathbf{v}_{\theta_\mathbf{n},t}(\mathbf{x}_{1:t-1})$
- **Both noisy observations and DVAE latent variables (NOLV):** $\mathbf{v}_{\theta_\mathbf{n},t} = \mathbf{v}_{\theta_\mathbf{n},t}(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$

Three training configurations

- **Unsupervised noise-agnostic (U-NA)**
- **Unsupervised noise-dependent (U-ND)**
- **U-NA fine-tuning after U-ND training (U-NDA)**



Experimental settings

Datasets

- VoiceBank-DEMAND (VB-DMD)
- WSJ0-QUT

Pre-processing

- STFT coefficients: 64-ms sine window (1,024 samples) and 75%-overlap (256-sample shift)

Baseline models

- Supervised methods: Open-Unmix (UMX), MetricGAN+, CDiffuSE, SGMSE+
- Unsupervised methods: MetricGAN-U, NyTT, RVAE-VEM

Evaluation metrics

- Enhancement performance: SI-SDR, PESQ (in [-0.5, 4.5]), ESTOI (in [0, 1])
- Computational efficiency: RTF

Experimental results

Table 1: Speech enhancement results.

Dataset	Training configuration	Model	SI-SDR ↑	PESQ_{MOS} ↑	ESTOI ↑
WSJ0-QUT	-	Noisy mixture	-2.6	1.83	0.50
	U-NA	RVAE-LV	5.4	2.31	**0.65**
		RVAE-NO	**6.0**	**2.33**	**0.65**
		RVAE-NOLV	5.5	2.31	**0.65**
	U-ND	RVAE-LV	**5.3**	**2.25**	**0.60**
		RVAE-NO	3.7	2.11	0.58
		RVAE-NOLV	4.9	2.11	**0.60**
	U-NDA	RVAE-LV	**6.2**	**2.38**	0.62
		RVAE-NO	5.8	2.31	**0.63**
		RVAE-NOLV	**6.2**	2.29	0.62
VB-DMD	Noisy mixture	-	8.4	3.02	0.79
	U-NA	RVAE-LV	**17.5**	3.23	**0.82**
		RVAE-NO	17.3	**3.25**	**0.82**
		RVAE-NOLV	**17.5**	**3.25**	**0.82**
	U-ND	RVAE-LV	**17.4**	**3.24**	**0.81**
		RVAE-NO	16.7	3.03	0.79
		RVAE-NOLV	16.9	3.04	0.79
	U-NDA	RVAE-LV	**17.8**	**3.22**	**0.81**
		RVAE-NO	17.2	3.06	0.80
		RVAE-NOLV	17.4	3.17	**0.81**

RVAE-LV: $\mathbf{v}_{\theta_n,t} = \mathbf{v}_{\theta_n,t}(\mathbf{z}_{1:T})$ RVAE-NO: $\mathbf{v}_{\theta_n,t} = \mathbf{v}_{\theta_n,t}(\mathbf{x}_{1:t-1})$ RVAE-NOLV: $\mathbf{v}_{\theta_n,t} = \mathbf{v}_{\theta_n,t}(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$

Speech enhancement examples

Model	Training configuration	Noisy speech	Reconstructed clean speech
RVAE-LV	U-NA		
	U-ND		
	U-NDA		

Comparison with the baselines

Table 2: Speech enhancement results. The baselines scores are taken from the corresponding papers. The best scores are in bold and the second best scores are underlined.

Dataset	Model	Supervision	SI-SDR ↑	$PESQ_{MOS}$ ↑	ESTOI ↑
WSJ0-QUT	Noisy mixture	-	-2.6	1.83	0.50
	UMX	Supervised	5.7	2.16	0.63
	MetricGAN+	Supervised	3.6	**2.83**	0.60
	RVAE-VEM	U-NA	5.8	2.27	0.62
	RVAE-LV	U-NA	5.4	2.31	**0.65**
		U-ND	5.3	2.25	0.60
		U-NDA	**6.2**	2.38	0.62
VB-DMD	Noisy mixture	-	8.4	3.02	0.79
	UMX	Supervised	14.0	3.18	0.83
	MetricGAN+	Supervised	8.5	**3.59**	0.83
	CDiffuSE	Supervised	12.6	-	0.79
	SGMSE+	Supervised	17.3	-	**0.87**
	NyTT Xtra	U-ND	17.7	-	-
	MetricGAN-U	U-ND	8.2	3.20	0.77
	RVAE-VEM	U-NA	17.1	3.23	0.81
	RVAE-LV	U-NA	17.5	3.23	0.82
		U-ND	17.4	3.24	0.81
		U-NDA	**17.8**	3.22	0.81

Inference computation time

Table 3: Inference computation time measured by the average real-time factor (RTF).

Dataset	Training configuration	Model	# Iteration	RTF
WSJ0-QUT	U-NA	RVAE-VEM	300	27.91
	U-NA	RVAE-LV	1000	89.42
		RVAE-NO	1000	89.34
		RVAE-NOLV	1000	90.98
	U-ND	RVAE-LV	0	**0.02**
		RVAE-NO	0	**0.02**
		RVAE-NOLV	0	**0.02**
	U-NDA	RVAE-LV	190	17.42
		RVAE-NO	500	45.54
		RVAE-NOLV	500	45.92
VB-DMD	SGMSE+	Supervised	-	3.39
	U-NA	RVAE-LV	900	81.62
		RVAE-NO	400	36.79
		RVAE-NOLV	800	73.24
	U-ND	RVAE-LV	0	**0.02**
		RVAE-NO	0	**0.02**
		RVAE-NOLV	0	**0.02**
	U-NDA	RVAE-LV	25	2.32
		RVAE-NO	25	2.13
		RVAE-NOLV	95	8.84

The real-time factor (RTF) is the time required to process 1 second of audio.

Conclusion

- We propose a new unsupervised speech enhancement model that uses a DDGM for both speech and noise.
- We tested three different dependencies for the noise model (NO, NOLV, LV), as well as three ‘training/testing’ configurations (U-NA, U-ND and U-NDA).
- Experimental results show that our model achieves comparable performance with the supervised and unsupervised baselines.
- In the ND configuration, our model provides a very fast inference process.