Speech Modeling with a Hierarchical Transformer Dynamical VAE

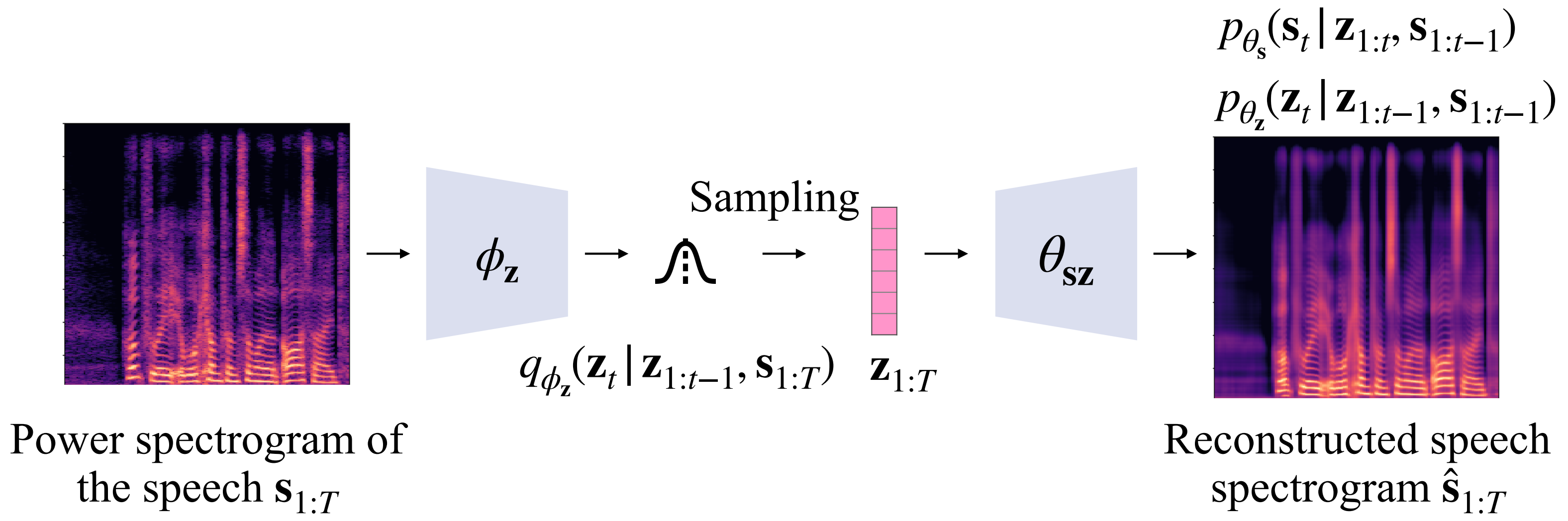
Xiaoyu Lin[1], Xiaoyu Bie[1], Simon Leglaive[2], Laurent Girin[3], Xavier Alameda-Pineda[1]

[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

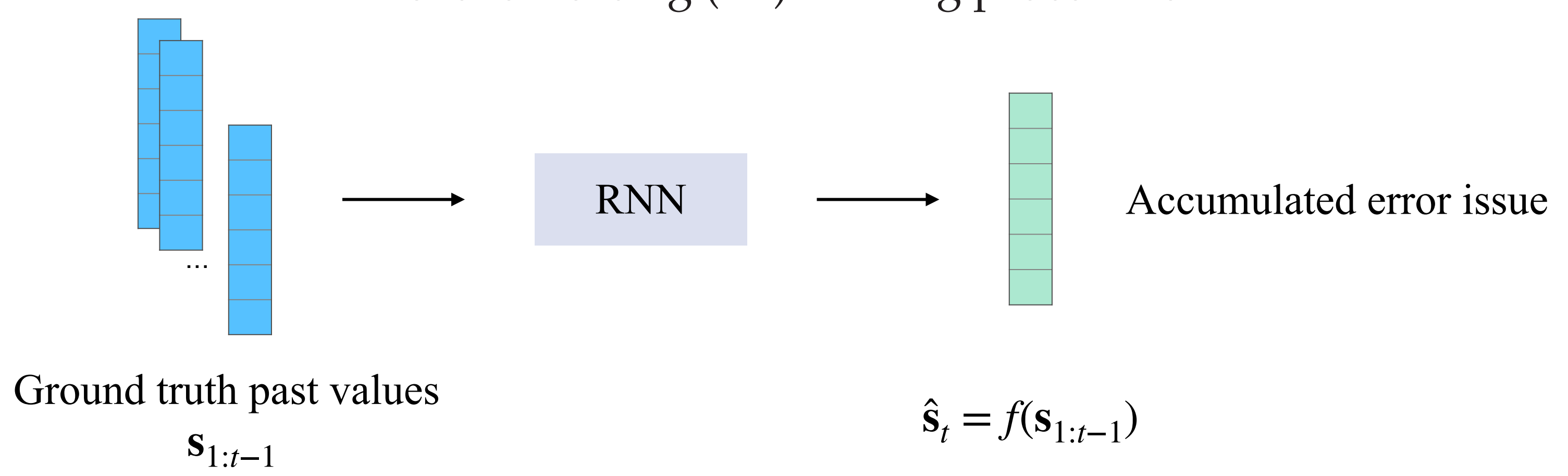
Context and motivation

Speech modeling with DVAEs

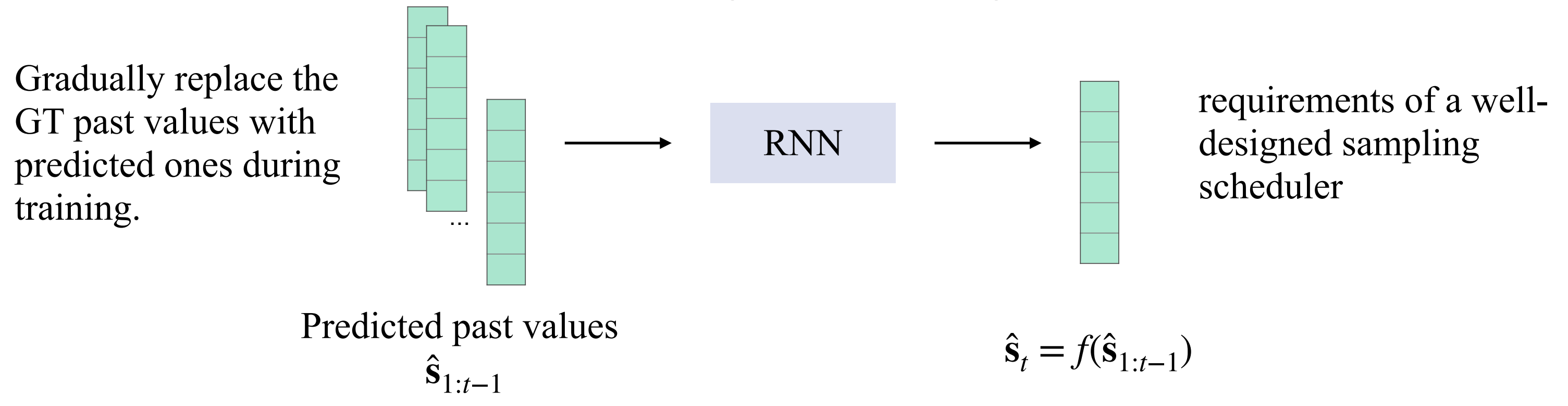

$p_{\theta_s}(\mathbf{s}_t|\mathbf{z}_{1:t}, \mathbf{s}_{1:t-1})$
$p_{\theta_z}(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{s}_{1:t-1})$
$\phi_{\mathbf{z}}$
Sampling
$\theta_{\mathbf{sz}}$
$q_{\phi_z}(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{s}_{1:T})$
$\mathbf{z}_{1:T}$
Power spectrogram of the speech $\mathbf{s}_{1:T}$
Reconstructed speech spectrogram $\hat{\mathbf{s}}_{1:T}$

RNN-based auto-regressive (AR) model training issues

Teacher forcing (TF) training procedure


...
RNN
Accumulated error issue
Ground truth past values $\mathbf{s}_{1:t-1}$
$\hat{\mathbf{s}}_t = f(\mathbf{s}_{1:t-1})$

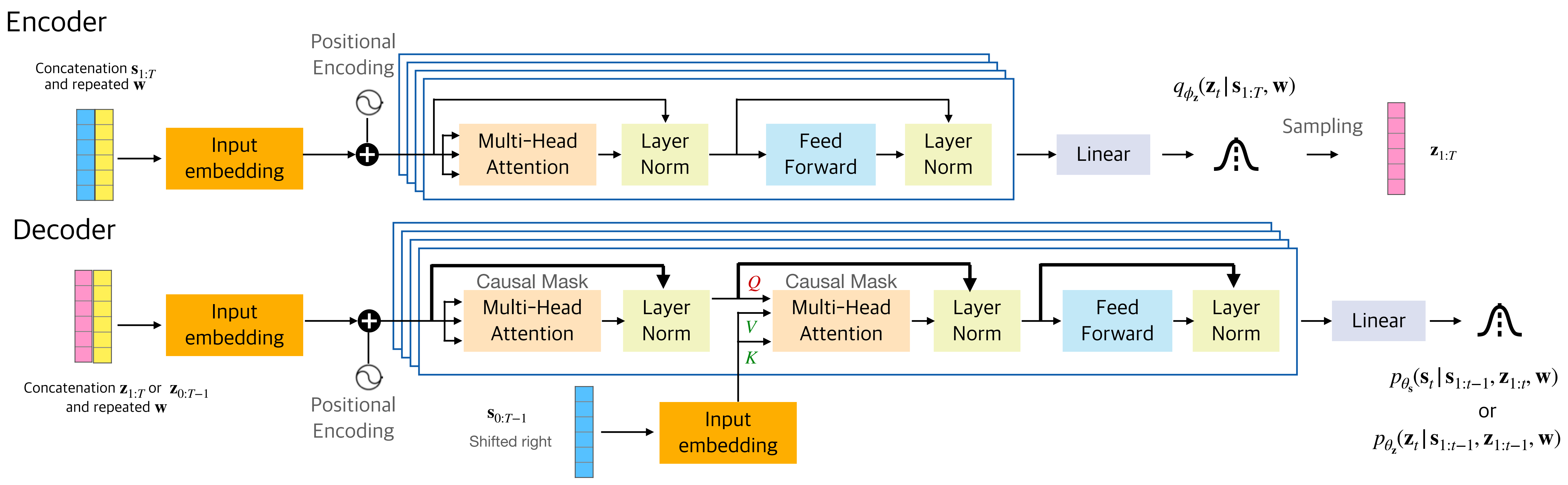
Scheduled sampling (SS) training procedure


Gradually replace the GT past values with predicted ones during training.
...
RNN
requirements of a well-designed sampling scheduler
Predicted past values $\hat{\mathbf{s}}_{1:t-1}$
$\hat{\mathbf{s}}_t = f(\hat{\mathbf{s}}_{1:t-1})$

Contributions

- Adapt the HiT-DVAE model to speech modeling, which was originally proposed for human pose generation.
- Propose the LigHT-DVAE model (share the parameters of the decoders), which reduces the model parameters of about 20% without degrading model performance.
- Investigate the HiT-DVAE and LigHT-DVAE model structures and explain the reason why the models are robust to the teacher-forcing training procedure.
- Investigate the generation ability of the HiT-DVAE and LigHT-DVAE models and compare them to the other DVAE models.

LigHT-DVAE model architecture


Encoder
Concatenation $\mathbf{s}_{1:T}$ and repeated $\mathbf{w}$
Input embedding
Positional Encoding
Multi-Head Attention
Layer Norm
Feed Forward
Layer Norm
Linear
$q_{\phi_z}(\mathbf{z}_t|\mathbf{s}_{1:T}, \mathbf{w})$
Sampling
$\mathbf{z}_{1:T}$
Decoder
Concatenation $\mathbf{z}_{1:T}$ or $\mathbf{z}_{0:T-1}$ and repeated $\mathbf{w}$
Input embedding
Positional Encoding
Causal Mask
Multi-Head Attention
Layer Norm
Q
V
K
Causal Mask
Multi-Head Attention
Layer Norm
Feed Forward
Layer Norm
Linear
$\mathbf{s}_{0:T-1}$
Shifted right
Input embedding
$p_{\theta_s}(\mathbf{s}_t|\mathbf{s}_{1:t-1}, \mathbf{z}_{1:t}, \mathbf{w})$
or
$p_{\theta_z}(\mathbf{z}_t|\mathbf{s}_{1:t-1}, \mathbf{z}_{1:t-1}, \mathbf{w})$

The model is trained by maximizing the Evidence Lower BOund (ELBO):

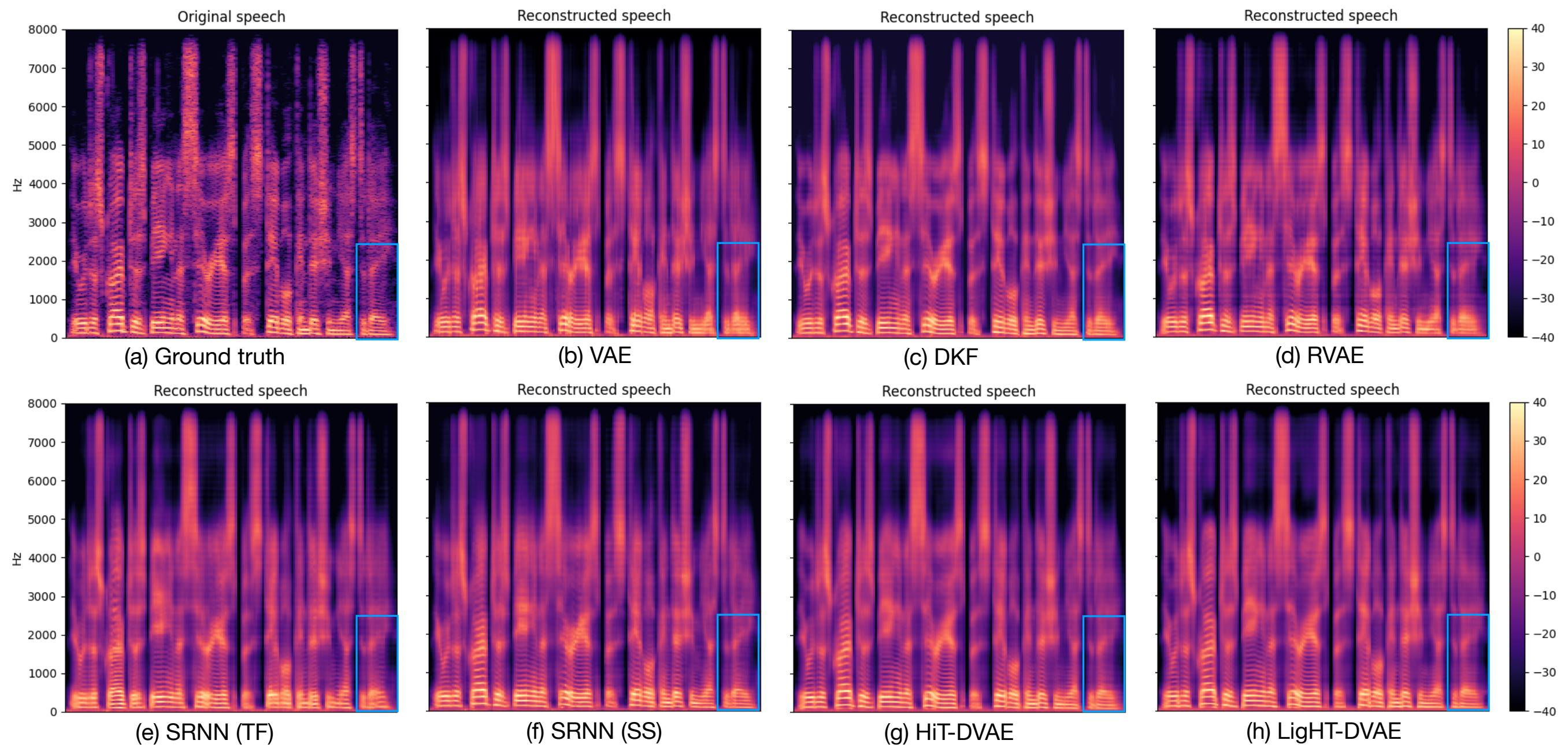
$$\mathcal{L}(\theta, \phi; \mathbf{s}_{1:T}) = -\underbrace{D_{\mathrm{KL}}(q_{\phi_{\mathbf{w}}}(\mathbf{w}|\mathbf{s}_{1:T})p_{\theta_{\mathbf{w}}}(\mathbf{w}))}_{\text{Regularization term for } \mathbf{w}} - \sum_{t=1}^{T} \mathbb{E}_{q_{\phi_{\mathbf{z}}} q_{\phi_{\mathbf{w}}}} \Big[\underbrace{d_{\mathrm{IS}}(|\mathbf{s}_t|^2, \mathbf{v}_{\theta_{\mathbf{s}},t})}_{\text{Reconstruction term}} + \underbrace{D_{\mathrm{KL}}(q_{\phi_{\mathbf{z}}}(\mathbf{z}_t|\mathbf{s}_{1:T}, \mathbf{w}) \parallel p_{\theta_{\mathbf{z}}}(\mathbf{z}_t|\mathbf{s}_{1:t-1}, \mathbf{z}_{1:t-1}, \mathbf{w}))}_{\text{Regularization term for } \mathbf{z}} \Big]$$

Speech Analysis-Resynthesis results

Speech analysis-resynthesis results.

Dataset	Model	RMSE ↓	SI-SDR ↑	PESQ ↑	ESTOI ↑
WSJ0	VAE	0.040	7.4	3.28	0.88
	DKF	0.037	8.3	3.51	**0.91**
	RVAE	0.034	8.9	3.53	**0.91**
	SRNN (SS)	0.036	8.7	**3.57**	**0.91**
	SRNN (TF)	0.061	2.6	2.53	0.76
	HiT-DVAE	0.031	10.0	3.52	**0.91**
	LigHT-DVAE	**0.030**	**10.1**	3.55	**0.91**
VB	VAE	0.052	8.4	3.24	0.89
	DKF	0.048	9.3	3.44	0.91
	RVAE	0.050	8.9	3.39	0.90
	SRNN (SS)	0.044	10.1	3.42	0.91
	SRNN (TF)	0.102	-0.1	2.15	0.75
	HiT-DVAE	0.039	11.4	**3.60**	**0.93**
	LigHT-DVAE	**0.038**	**11.6**	3.58	**0.93**

Speech spectrograms of analysis-resynthesis examples.



(a) Ground truth (b) VAE (c) DKF (d) RVAE

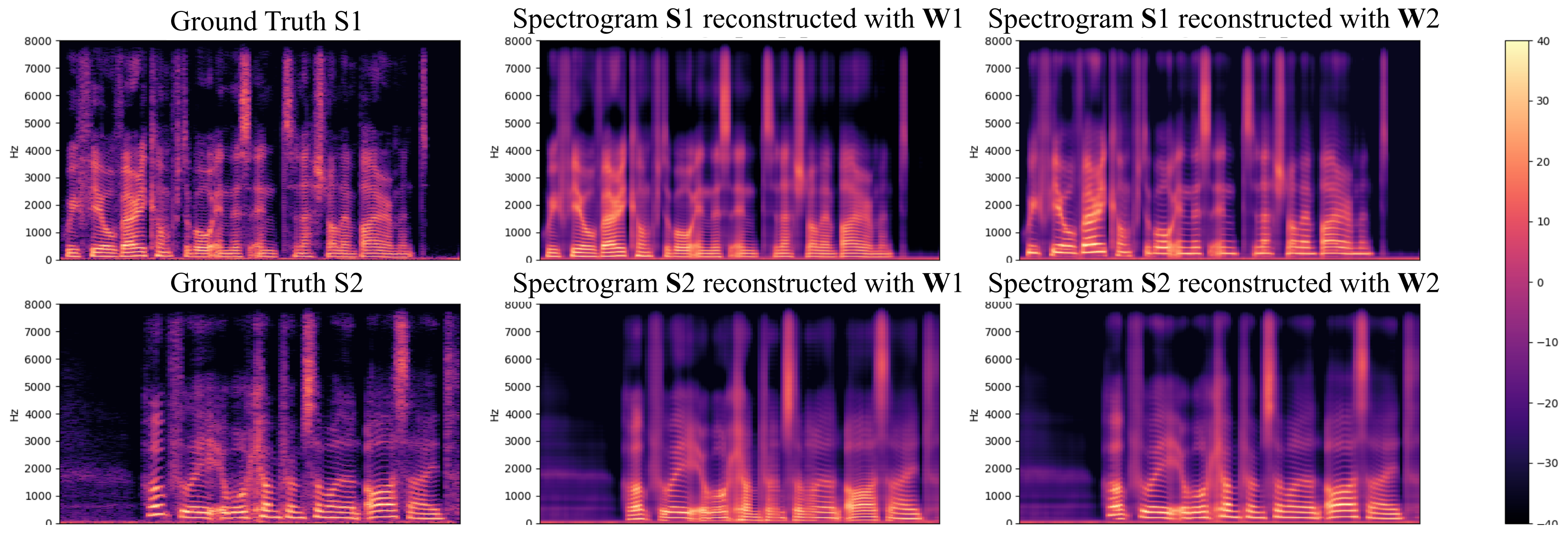
(e) SRNN (TF) (f) SRNN (SS) (g) HiT-DVAE (h) LigHT-DVAE

Investigation on the model structures.

Test $\mathbf{s}_{1:t-1}$	Model	RMSE ↓	SI-SDR ↑	PESQ ↑	ESTOI ↑
GEN	HiT-DVAE	0.039	11.4	3.60	0.93
	HiT-DVAE-Inv-s	0.079	3.8	2.61	0.75
	HiT-DVAE-Inv-s-NR	0.067	5.8	2.68	0.78
	LigHT-DVAE	0.038	11.6	3.58	0.93
	LigHT-DVAE-Inv-s	0.079	3.9	2.58	0.75
	LigHT-DVAE-Inv-s-NR	0.068	5.7	2.63	0.78
GT	HiT-DVAE	0.038	11.5	3.60	0.93
	HiT-DVAE-Inv-s	0.038	11.4	3.32	0.90
	HiT-DVAE-Inv-s-NR	0.067	5.8	2.68	0.78
	LigHT-DVAE	0.038	11.7	3.59	0.93
	LigHT-DVAE-Inv-s	0.040	10.9	3.29	0.89
	LigHT-DVAE-Inv-s-NR	0.068	5.7	2.63	0.78

Investigation on w

Reconstructed spectrograms by exchanging **w**.

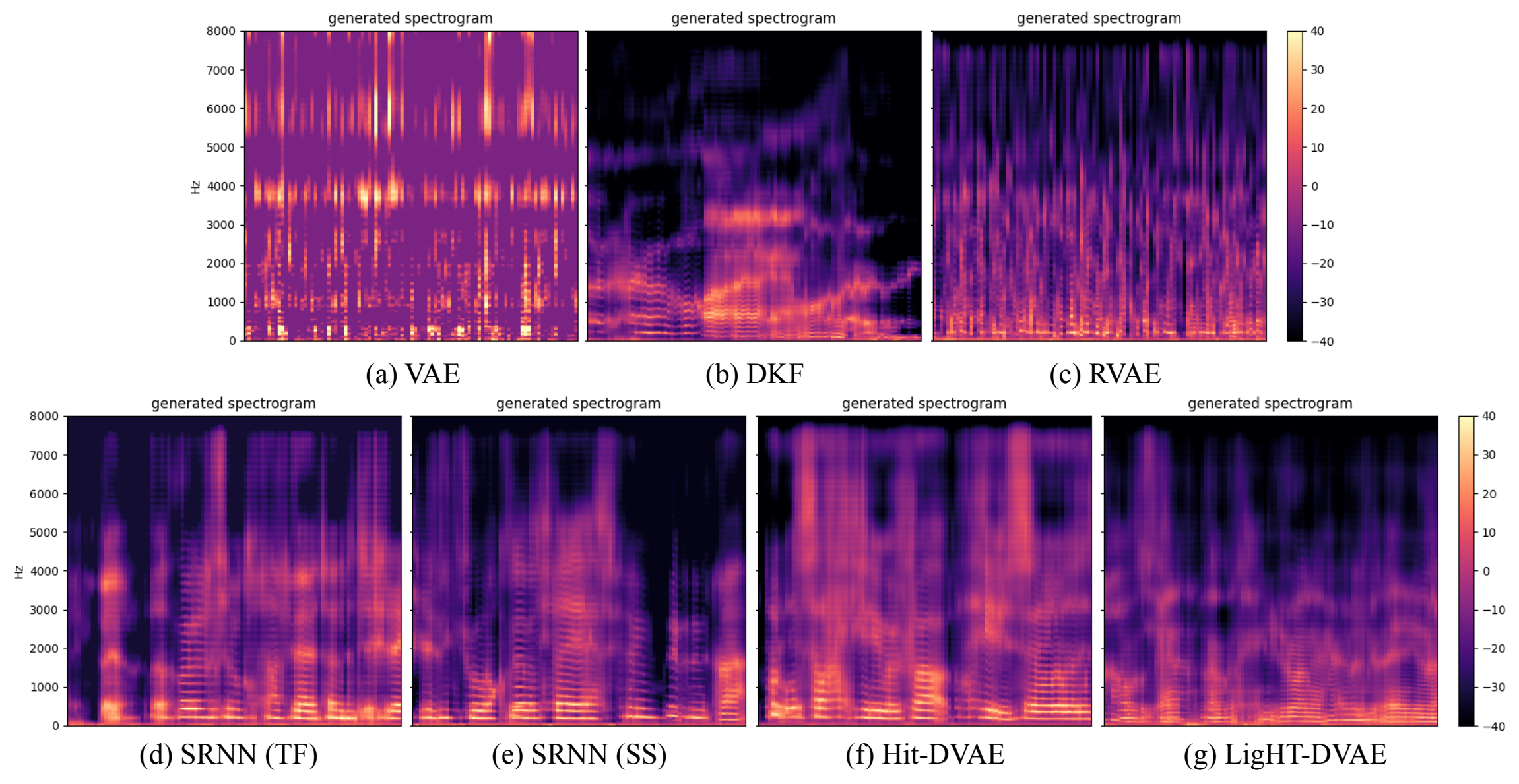


Generation Results

Speech spectrograms generation results.

Model	FDSD ↓
VAE	70.92 ± 0.44
DKF	32.78 ± 0.28
RVAE	45.75 ± 0.11
SRNN (SS)	25.28 ± 0.19
SRNN (TF)	25.53 ± 0.13
HiT-DVAE	**22.50 ± 0.26**
LigHT-DVAE	29.22 ± 0.26
VB Test (exact phase)	4.11 ± 0.14
VB Test (Griffin-Lim)	4.11 ± 0.15

Speech spectrograms generation examples.



(a) VAE (b) DKF (c) RVAE

(d) SRNN (TF) (e) SRNN (SS) (f) Hit-DVAE (g) LigHT-DVAE