Firstly, we thank all reviewers for the helpful comments and suggestions.

To Reviewer 2:

$q(z_t|z_{t-1}, \mathbf{x}_{<t})$ in Eq (2) is a typo. The correct one should be $q(z_t|z_{t-1}, \mathbf{x})$, which is derived from the autoregressive factorization of $q(\mathbf{z}|\mathbf{x})$, $q(\mathbf{z}|\mathbf{x}) = \prod_{t=1}^{T} q(z_t|z_{t-1}, \mathbf{x})$. Thanks for spotting and pointing out the typo.

In the information leaking experiment, each multivariate one-step observation x_t is split into two vectors x_t^a and x_t^b. Computational, we first summarize the historical information before time step t with an RNN and denote it as $h_t = f(\mathbf{x}_{<t}) \coloneqq \text{RNN}(x_{t-1}, h_{t-1})$. Then, $p(x_t \mid \mathbf{x}_{<t}) = p(x_t^a \mid \mathbf{x}_{<t})p(x_t^b \mid x_t^a, \mathbf{x}_{<t})$ are parameterized as two multivariate Gaussians: $p(x_t^a \mid \mathbf{x}_{<t}) \coloneqq \mathcal{N}(x_t^a; \mu_a(h_t), I\sigma_a^2(h_t))$ and $p(x_t^b \mid x_t^a, \mathbf{x}_{<t}) \coloneqq \mathcal{N}(x_t^b; \mu_b(h_t, x_t^a), I\sigma_b^2(h_t, x_t^a))$, where μ_a, σ_a and μ_b, σ_b are all trainable MLPs that output the vector-valued mean and (diagonal) variance for the corresponding distributions. Hence, x_t^a is treated as a vector of dimension $|x_t^a|$ instead of a sequence when fed into μ_b, σ_b.

In our experiments, by decreasing L, the gap between F-SRNN and F-RNN decreases, and gradually F-RNN outperforms F-SRNN. However, by using the RNN-hier architecture in our paper, the deterministic RNN model outperforms the SRNN model in the settings with any L value.

We will add citations in Table 4. We haven't conducted experiments in language modeling and image density estimation tasks. But from existing publications, the state-of-the-art results of these tasks are produced by auto-regressive style models.

To Reviewer 3:

Thanks for your suggestion on comparing the running time of different models. We will include this part of the results in the revised version. The running times of training models for 40k updating steps on TIMIT are summarized in Table 1.

Input Length	8000							1000
Model Name	F-RNN	F-SRNN	δ-RNN	RNN-hier	SRNN-hier	RNN-flat	SRNN-flat	RNN-hier
Training Time	0.54h	0.94h	0.90h	9.92h	12.52h	37.48h	42.26h	1.7h
Log-Likelihood	32,745	69,296	66,453	109,641	107,912	117,721	109,284	101,713

Table 1: Training time comparison between various models.

Admittedly, modeling the intra-step correlation would require extra computation time. Hence, this leads to a trade-off between quality and speed. Ideally, latent-variable models would provide a solution close to the sweet point of this trade-off. However, in our experiment, we find a simple hierarchical auto-regressive model trained with a shorter input length could already achieve significantly better performance with a comparable computation time (RNN-hier vs. F-SRNN in Table 1). We will add this discussion in the revised version.

Finally, we would like to emphasize that the goal of this work is to perform a fair and informative reexamination of recurrent stochastic models rather than downplay any model. Based on our analysis and empirical evidence, we hope to (1) correct the previous misleading conclusion that SRNN can already achieve better results compared with deterministic RNN in the sequential density estimation, (2) provide a more realistic benchmark with SOTA baselines for speech density estimation and encourage future researchers to perform a more meaningful model comparison, (3) offer some informative analysis and understanding of what SRNN is actually doing in practice. Overall, we may still have a long way to go to really fulfill the theoretical advantage of stochastic sequential models.

To Reviewer 4: We think the massive improvement provided by the auto-regressive model (including column 2 and other columns) shows that the performance of the deterministic model is heavily underestimated in the previous biased experiment setting.

We are not entirely sure about the motivation of the multi-frame setting. One possibility is to simulate the case of modeling natural multi-variate sequences such the midi music. The computation speed could be another consideration because the sequence length of speech data is much longer than language and image data, whose sample rate is 16k per second.

We have not conducted in-depth research on different sample rates yet. According to popular speech synthesis papers, WaveNet uses 16k sample rate and DeepVoice uses 16k and 48k.