Supplementary Materials:
Deep Graph Random Process for Relational-Thinking-Based Speech Recognition

1 Proof of Theorem 1

Theorem 1. *Let $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with $\mu < 1/2$, and let $\mathcal{B}(n, \lambda)$ denotes a Binomial distribution with $n \to +\infty$ and $\lambda \to 0$, where n is increasing while λ is decreasing. There exists a real constant m such that if $m = n\lambda$ and if we define:*

$$\begin{aligned} f_1(x) &= \mathrm{KL}(\mathcal{N}(x, x(1-x))||\mathcal{N}(\mu, \sigma^2)) \\ f_2(x) &= \mathrm{KL}(\mathcal{N}(x, x(1-x))||\mathcal{N}(n\lambda, n\lambda(1-\lambda)) \\ f_2^* &= \min_x f_2(x),\ where\ x \in (0,1) \end{aligned}$$

we have that: $f_1(x)$ attains its minimum on the interval $(0, 1)$ and $f_2(x) - f_2^*$ is bounded on the interval $(0, \sqrt{2}/2 - 1/2)$, with:

$$x = m = \frac{1 + l - \sqrt{1 + l^2}}{2},\ where\ l = \frac{2\sigma^2}{1 - 2\mu}$$

Proof. The derivative of the function $f_1(x)$ over x can be written as:

$$f_1'(x) = x^2 - (1 + \frac{2\sigma^2}{1 - 2\mu})x + \frac{\sigma^2}{1 - 2\mu}$$

We set it as 0 and solve for x, giving

$$x = \begin{cases} \frac{1+l-\sqrt{1+l^2}}{2} & \text{if } \mu < 1/2 \\ \frac{1+l+\sqrt{1+l^2}}{2} & \text{if } \mu > 1/2 \end{cases},\ where\ l = \frac{2\sigma^2}{1 - 2\mu} \tag{1}$$

Let $x = n\lambda$, the function $f_2(x)$ can be written as:

$$f_2(n\lambda) = \sqrt{\frac{1 - n\lambda}{1 - \lambda}} + \frac{1 - \lambda}{2(1 - n\lambda)} - 1/2$$

Let $g(n\lambda) = \lim_{\lambda \to 0} f_2(n\lambda)$, we have

$$g(n\lambda) = \sqrt{1 - n\lambda} + \frac{1}{2(1 - n\lambda)} - 1/2$$

Let $z = \sqrt{1 - n\lambda}$, we have:

$$g(z) = z + 1/(2z^2) - 1/2$$

The derivative of function $g(z)$ over z can be written as:

$$g'(z) = 1 - 1/z^3$$

Given that $z \in (0,1)$, we have $g'(z) < 0$. Then $g(z)$ attains its minimum 1 when z approaches 1. Equivalently, $f_2(n\lambda)$ attains its minimum 1 when $n\lambda$ approaches 0.

Considering Eq.(1), we find that $n\lambda$ is bounded on $(0, 1/2)$ if $\mu < 1/2$,

We then calculate the difference between $f_2(n\lambda)$ and its minimum. It can be written as

$$\begin{aligned}\Delta f_2(n\lambda) &= \lim_{\lambda \to 0}[f_2(x) - f_2^*] \\ &= g(n\lambda) - 1 \\ &= \sqrt{1-n\lambda} + \frac{1}{2(1-n\lambda)} - 3/2\end{aligned}$$

Let $m = n\lambda$, the derivative of function $\Delta f_2(m)$ over m can be written as:

$$\Delta f_2'(m) = \frac{1-(1-m)^{3/2}}{2(1-m)^2} > 0$$

Then $\Delta f_2(m)$ is monotonically increasing over $(0, 1/2)$. Therefore $\Delta f_2(m)$ is bounded on $(0, \sqrt{2}/2 - 1/2)$ □

2 Proof of Theorem 2

Theorem 2. *Suppose we are given two Binomial distributions, $\mathcal{B}(n, \lambda)$ and $\mathcal{B}(n, \lambda^0)$ with $n \to +\infty$, $\lambda^0 \to 0$ and $\lambda \to 0$, where n is increasing while λ and λ^0 are decreasing. There exists a real constant m and another real constant $m^{(0)}$, such that if $m = n\lambda$ and $m^{(0)} = n\lambda^{(0)}$ and if $\lambda > \lambda^{(0)}$, we have:*

$$\mathrm{KL}(\mathcal{B}(n,\lambda)||\mathcal{B}(n,\lambda^0)) < m \log \frac{m}{m^{(0)}} + (1-m)\log\frac{1-m+m^2/2}{1-m^{(0)}+{m^{(0)}}^2/2}$$

Proof. Let $m = n\lambda$ and $m^{(0)} = n\lambda^{(0)}$, we have

$$\begin{aligned}\mathrm{KL}(\mathcal{B}(n,\lambda)||\mathcal{B}(n,\lambda^{(0)})) &= n\lambda \log\frac{\lambda}{\lambda^{(0)}} + n(1-\lambda)\log\frac{1-\lambda}{1-\lambda^{(0)}} \\ &= n\lambda \log\frac{n\lambda}{n\lambda^{(0)}} + n(1-\lambda)\log\frac{1-\lambda}{1-\lambda^{(0)}} \\ &= m \log\frac{m}{m^{(0)}} + n(1-\lambda)\log\frac{1-\lambda}{1-\lambda^{(0)}}\end{aligned} \tag{2}$$

We then take the right part,

$$g = n(1-\lambda)\log\frac{1-\lambda}{1-\lambda^{(0)}} = (1-\lambda)\log\frac{(1-\lambda)^n}{(1-\lambda^{(0)})^n}$$

By Taylor series' theorem with Lagrange remainder, g can be written as:

$$g = (1-\lambda)\log\frac{1-n\lambda+\frac{n(n-1)}{2}\lambda^2 + R_{j=2}(-\lambda)}{1-n\lambda^{(0)}+\frac{n(n-1)}{2}{\lambda^{(0)}}^2 + R_{j=2}(-\lambda^{(0)})}$$

There exists a $\theta \in (0,1)$ such that,

$$R_{j=2}(x) = \frac{x^3(n-2)(n-1)n(1+x\theta)^{n-3}}{6}$$

Given that $n \to +\infty$ and $x \in (-1, 1)$, we have $(R)'_{j=2}(x) > 0$. Therefore, $R_{j=2}(x)$ is monotonically increasing over $(-1, 1)$. Since $\lambda > \lambda^0$, we have

$$R_{j=2}(-\lambda) < R_{j=2}(-\lambda^{(0)}) \tag{3}$$

We then seek to prove:

$$k = \frac{1 - n\lambda + \frac{n(n-1)}{2}\lambda^2}{1 - n\lambda^{(0)} + \frac{n(n-1)}{2}{\lambda^{(0)}}^2} < 1$$

Let $f(x) = n(n-1)x^2/2 - nx + 1$, we have

$$k = \frac{f(\lambda)}{f(\lambda^{(0)})}$$

Here, $f(x)$ is an U-shaped parabola with axis $x = 1/(n-1)$. By theorem 1, we have $n\lambda < 1/2$, then we have $\lambda^0 < \lambda < 1/(n-1)$, then $f(x)$ is monotonically increasing over the support of λ^0 and λ, namely

$$f(\lambda) < f(\lambda^0) \tag{4}$$

With Eq.(3) and Eq.(4), g can be written as:

$$\begin{aligned} g &< (1-\lambda)\log\frac{1 - n\lambda + \frac{n(n-1)}{2}\lambda^2}{1 - n\lambda^{(0)} + \frac{n(n-1)}{2}{\lambda^{(0)}}^2} \\ &= (1-\lambda)\log\frac{1 - m + m^2/2 - n\lambda^2/2}{1 - m^{(0)} + {m^{(0)}}^2/2 - n{\lambda^{(0)}}^2/2} \end{aligned}$$

Similarly, let $h(x) = 1 - x + x^2/2$. It is an U-shaped parabola with axis $x = 1$ such that

$$-n\lambda^2/2 < -n{\lambda^{(0)}}^2/2$$

$$1 - m + m^2/2 < 1 - m^{(0)} + {m^{(0)}}^2/2$$

Then we have

$$\begin{aligned} g &< (1-\lambda)\log\frac{1 - m + m^2/2}{1 - m^{(0)} + {m^{(0)}}^2/2} \\ &< (1-n\lambda)\log\frac{1 - m + m^2/2}{1 - m^{(0)} + {m^{(0)}}^2/2} \\ &= (1-m)\log\frac{1 - m + m^2/2}{1 - m^{(0)} + {m^{(0)}}^2/2} \end{aligned} \tag{5}$$

Combining Eq.(2) and Eq.(5) concludes the proof. □

3 Test of Significance

The statistical significance test tool sc_stats from National Institute of Standards and Technology (NIST) is used to compare our RTN and the baseline VSRU on CHiME-2 HMM states classification task. The test results find a significant difference in performance between the RTN and the VSRU at the level of $p < 0.001$.

4 Table of detailed WER (%) on the CHiME-2 test set

We report the detailed WERs as a function of the SNR in CHiME-2 shown in Table 1. For all SNRs, the RTN outperforms other Baseline RNNs including LSTM, SRU by a large margin. It outperforms the state-of-the-art models including VSRU, RRN and RPPU for most SNRs. This suggests that incorporating the relational thinking into speech recognition lends itself to the model's robustness.

Table 1: Detailed WER (%) on the CHiME-2 test set.

Model	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
LSTM [Huang et al., 2019]	42.4	33.5	26.7	21.1	17.3	15.3
SRU [Huang et al., 2019]	42.5	34.0	26.2	22.2	17.4	15.1
RPPU [Huang et al., 2019]	39.9	31.1	24.9	20.3	16.0	**13.2**
Our SRU [Lei et al., 2017]	42.1	33	26.1	20.7	16.8	15.1
VSRU [Chung et al., 2015]	41.5	32.8	26.2	20.9	16.9	16.1
RRN [Palm et al., 2018]	40.2	32.1	25.9	20.2	16.2	14.0
RTN (Ours)	**39.0**	**30.4**	**25.4**	**19.4**	**15.5**	13.8

5 More examples of graphs generated by RTN

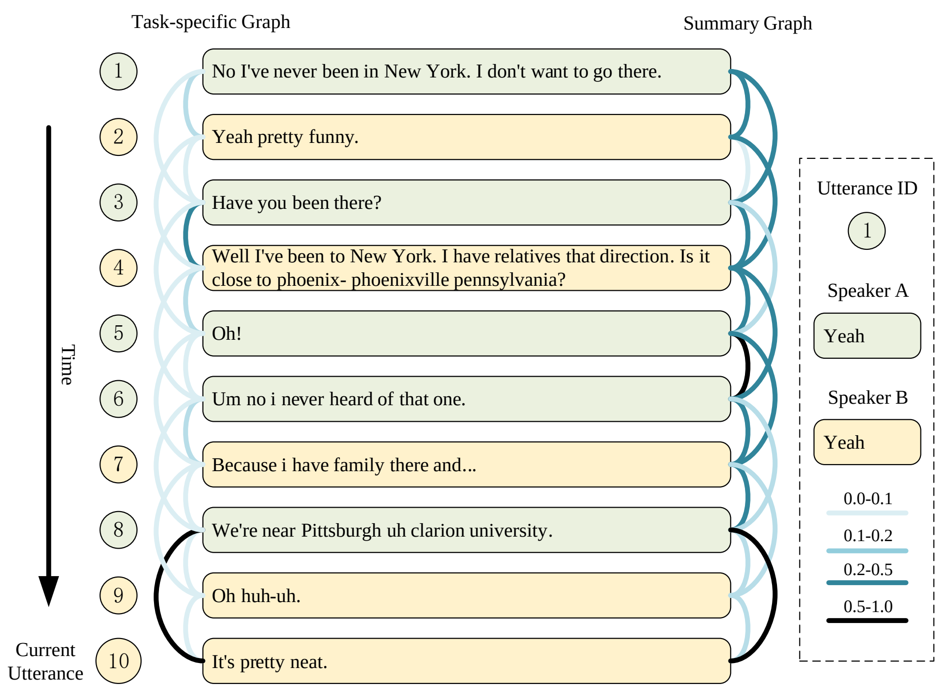

Figure 1: Example of graphs generated by RTN: ten sequential utterances from "sw02262-A_029098-029769" to "sw02262-B_031645-031828"

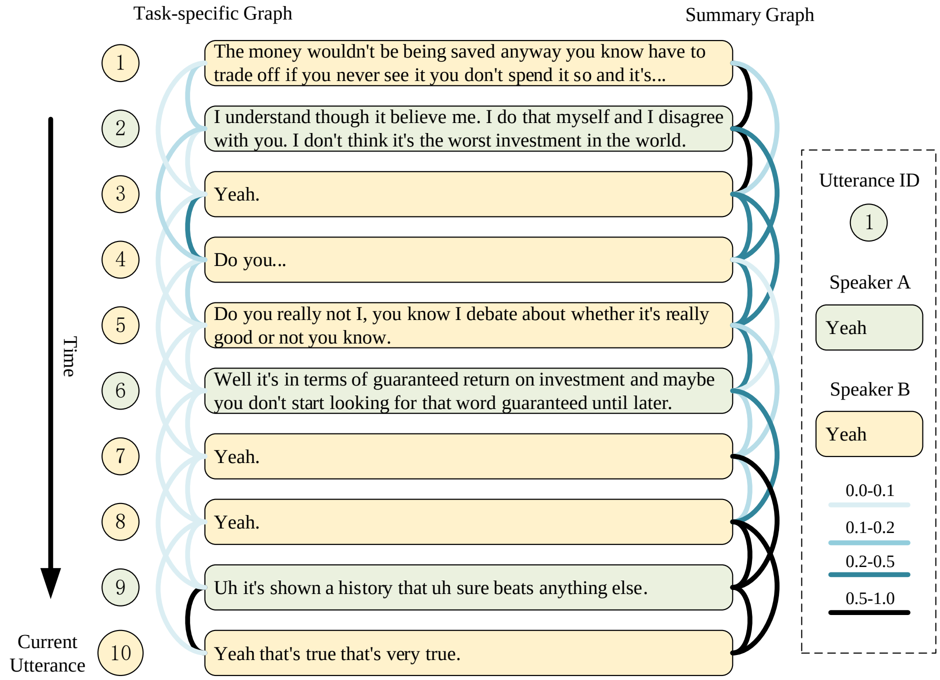

Figure 2: Example of graphs generated by RTN: ten sequential utterances from ”sw02062-B_019277-020062” to ”sw02062-B_022871-023232”

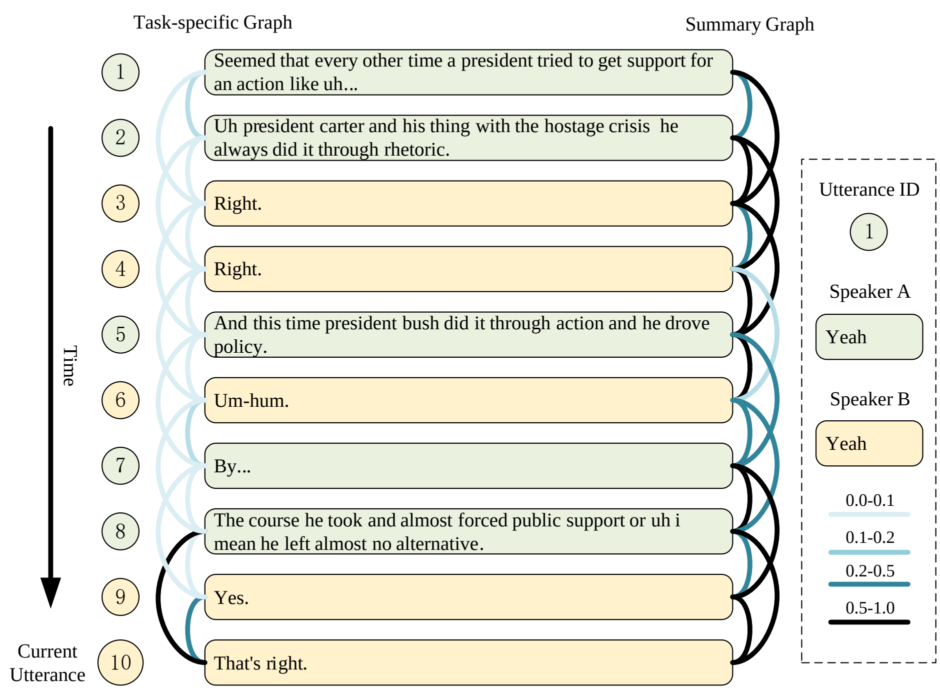

Figure 3: Example of graphs generated by RTN: ten sequential utterances from ”sw02130-A_002749-003357” to ”sw02130-B_005687-005840”

References

[Chung et al., 2015] Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. (2015). A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988.

[Huang et al., 2019] Huang, H., Wang, H., and Mak, B. (2019). Recurrent poisson process unit for speech recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6538–6545.

[Lei et al., 2017] Lei, T., Zhang, Y., Wang, S. I., Dai, H., and Artzi, Y. (2017). Simple recurrent units for highly parallelizable recurrence. *arXiv preprint arXiv:1709.02755*.

[Palm et al., 2018] Palm, R., Paquet, U., and Winther, O. (2018). Recurrent relational networks. In *Advances in Neural Information Processing Systems*, pages 3368–3378.