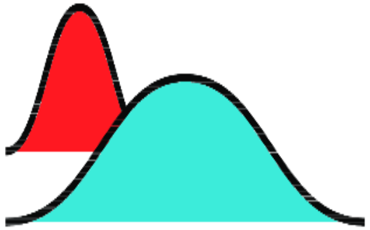
Deep Graph Random Process for Relational-Thinking-Based Speech Recognition

HENGGUAN HUANG,

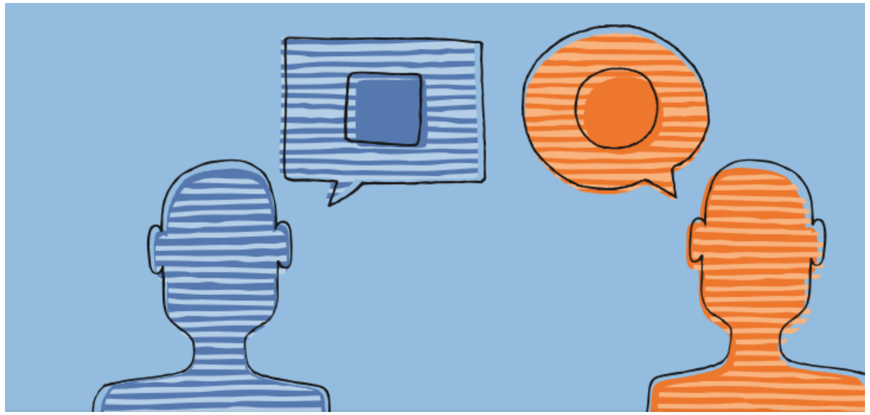
FUZHAO XUE, HAO WANG, YE WANG

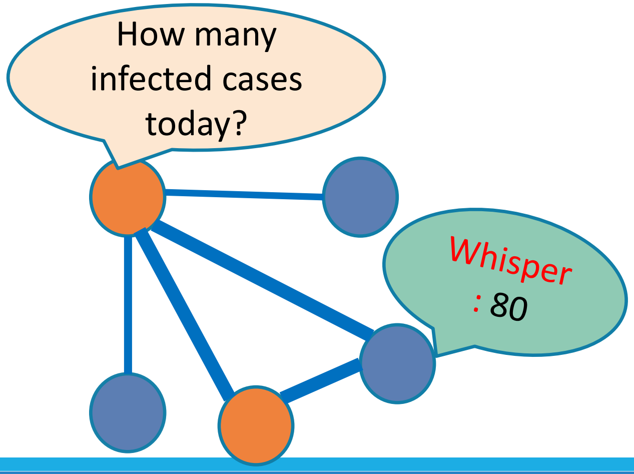
Conversational Speech Recognition



Bayesian Deep learning

Neurobiology:
Relational Thinking





Motivation: relational thinking

Motivation: relational thinking

A type of human learning process, in which people spontaneously perceive meaningful patterns from the surrounding world .

A relevant concept: percept

- Unconscious mental impressions while hearing, seeing…
- Relations between current sensory signals and prior knowledge

Patricia A Alexander. Relational thinking and relational reasoning: harnessing the power of patterning. NPJ science of learning, 1:16004, 2016

Motivation: Relational thinking

A type of human learning process, in which people spontaneously perceive meaningful patterns from the surrounding world .

Two-step procedure:

- **Step 1**: the generation of an infinite number of ***percepts***
- **Step 2**: these percepts are then combined and transformed into concept or idea

Largely unexplored in AI (**focus of this project**)

Patricia A Alexander. Relational thinking and relational reasoning: harnessing the power of patterning. NPJ science of learning, 1:16004, 2016

Overview

- **Our Goal**: relational thinking modeling and its application in acoustic modeling
- **Challenges** (if percepts are modelled as graphs):
  - Edges in the graph are not annotated/available (no relational labels)
  - Hard to optimize over an infinite number of graphs
- **Existing works:**
  - GNNs (e.g. GVAE) require input/output to have graph structure
  - Can not handle an infinite number of graphs
  - Current acoustic models (e.g. RNN-HMM, the model we works on) is limited in representing complex relationships

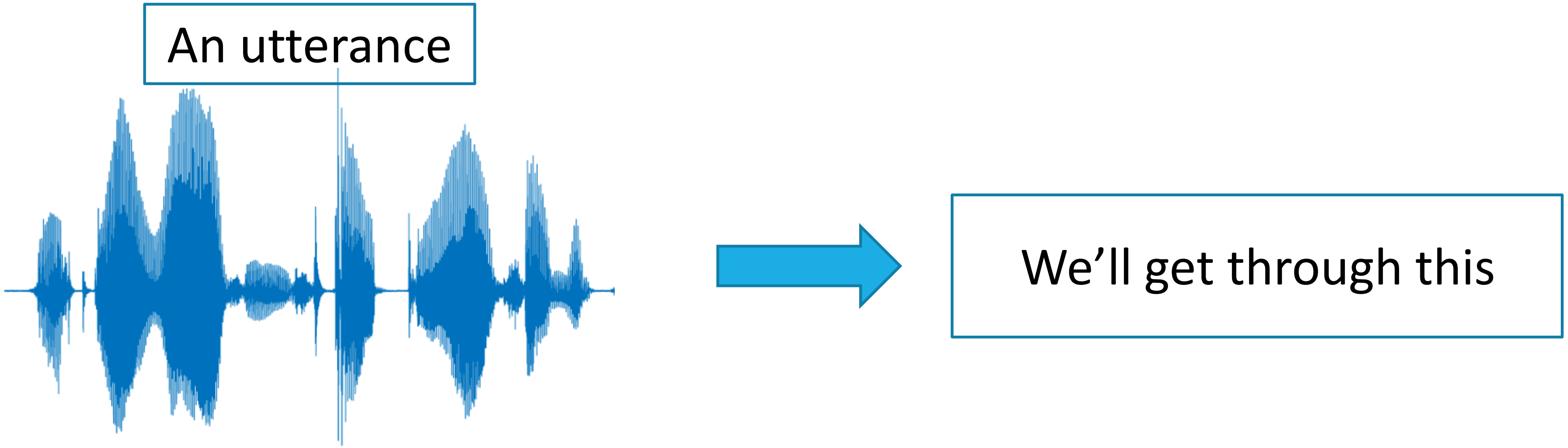
Overview

- **Our Solution**:
  - Build a type of random process that can simulate generation of an infinite number of percepts (graphs) called deep graph random process (DGP)
  - Provide a close-form solution for combining an infinite number of graphs (coupling of percepts)
  - Apply DGP for acoustic modelling (transformation of percetps)
  - Obtain an analytical ELBO for jointly training
- **Advantages**:
  - Relation labels is not required during training
  - Easy to apply for down-stream tasks, e.g. ASR
  - Computationally efficient and better performance

Machine speech recognition

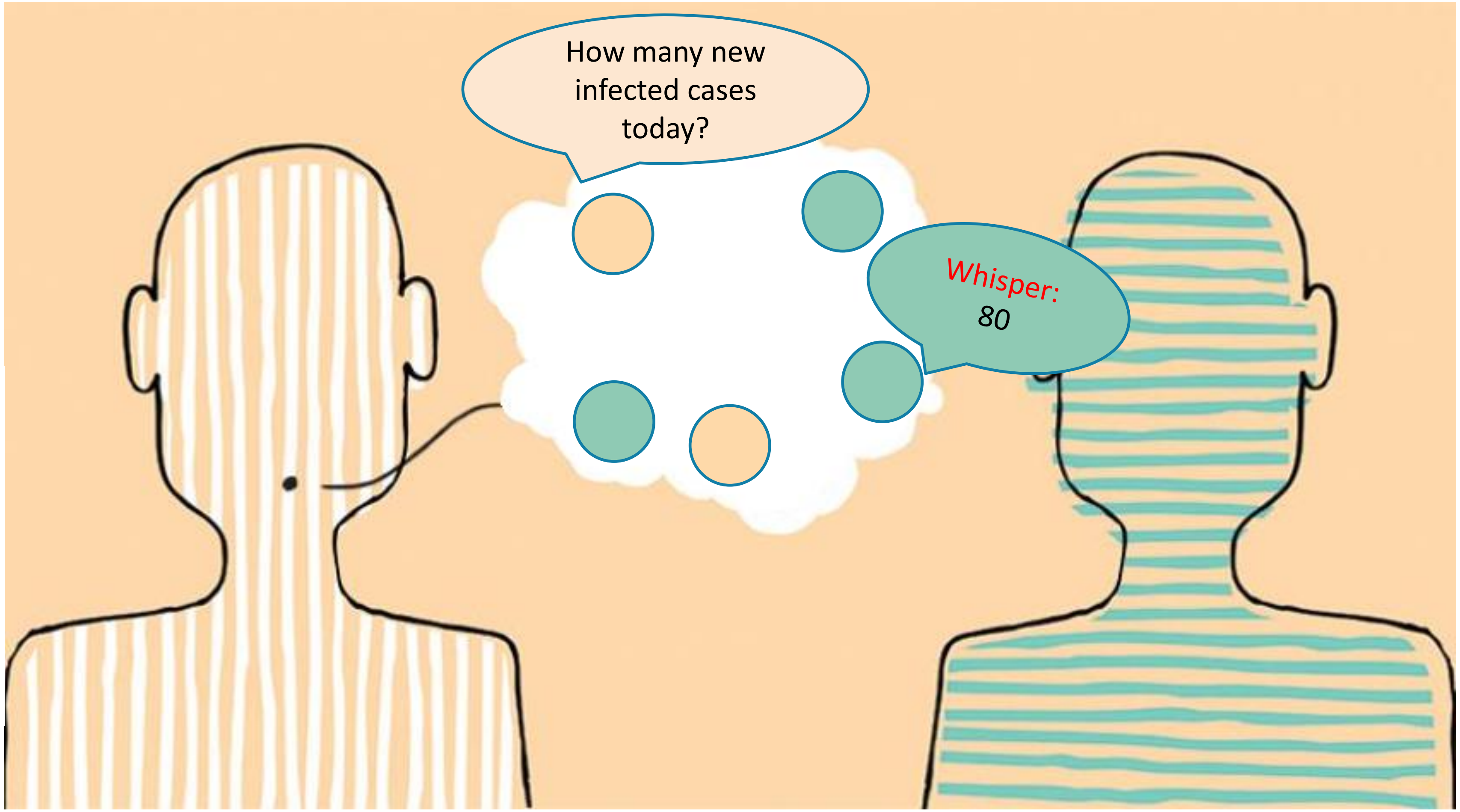
Speech-to-text transcription

- Transform audio into words

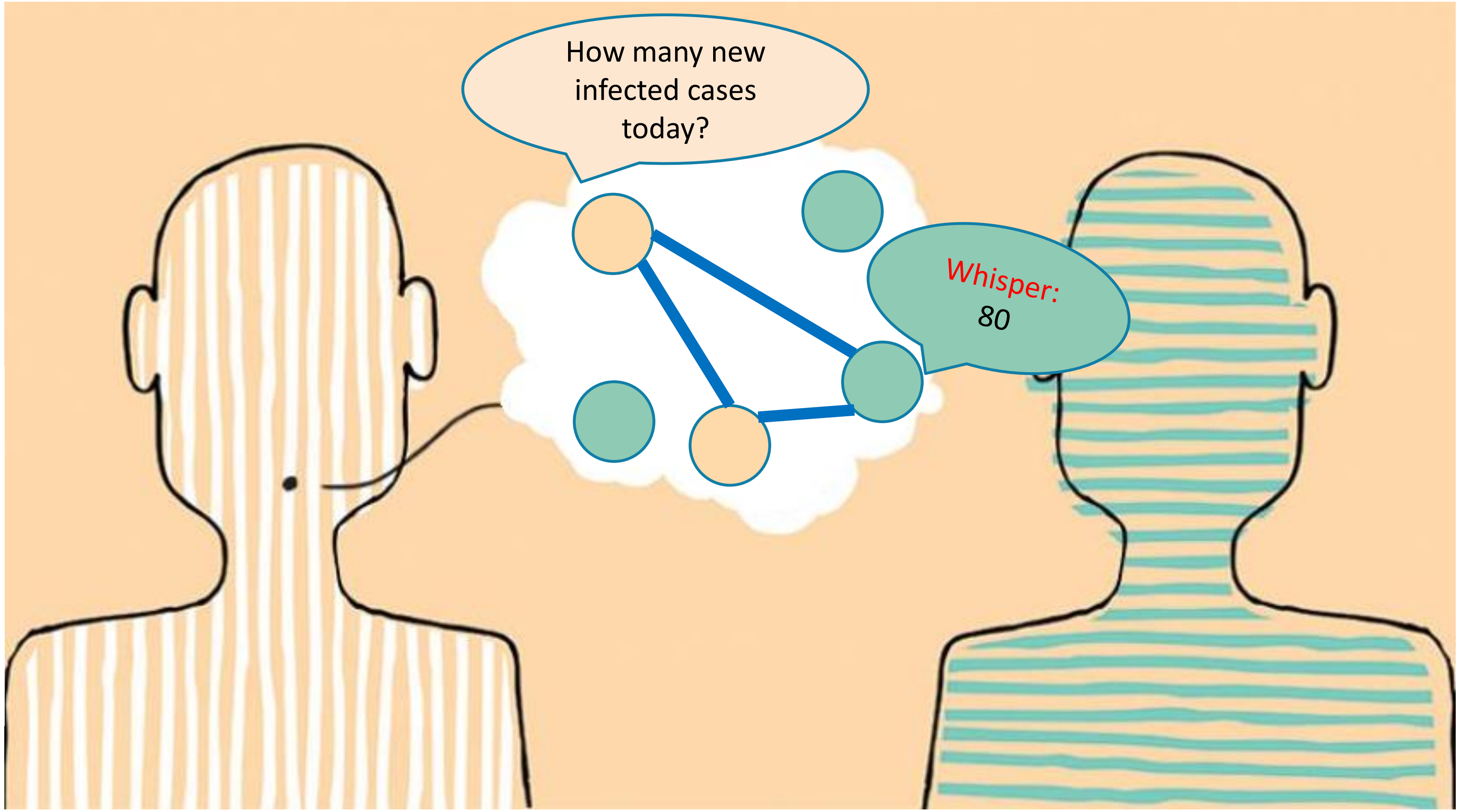


- Relational thinking process is ignored

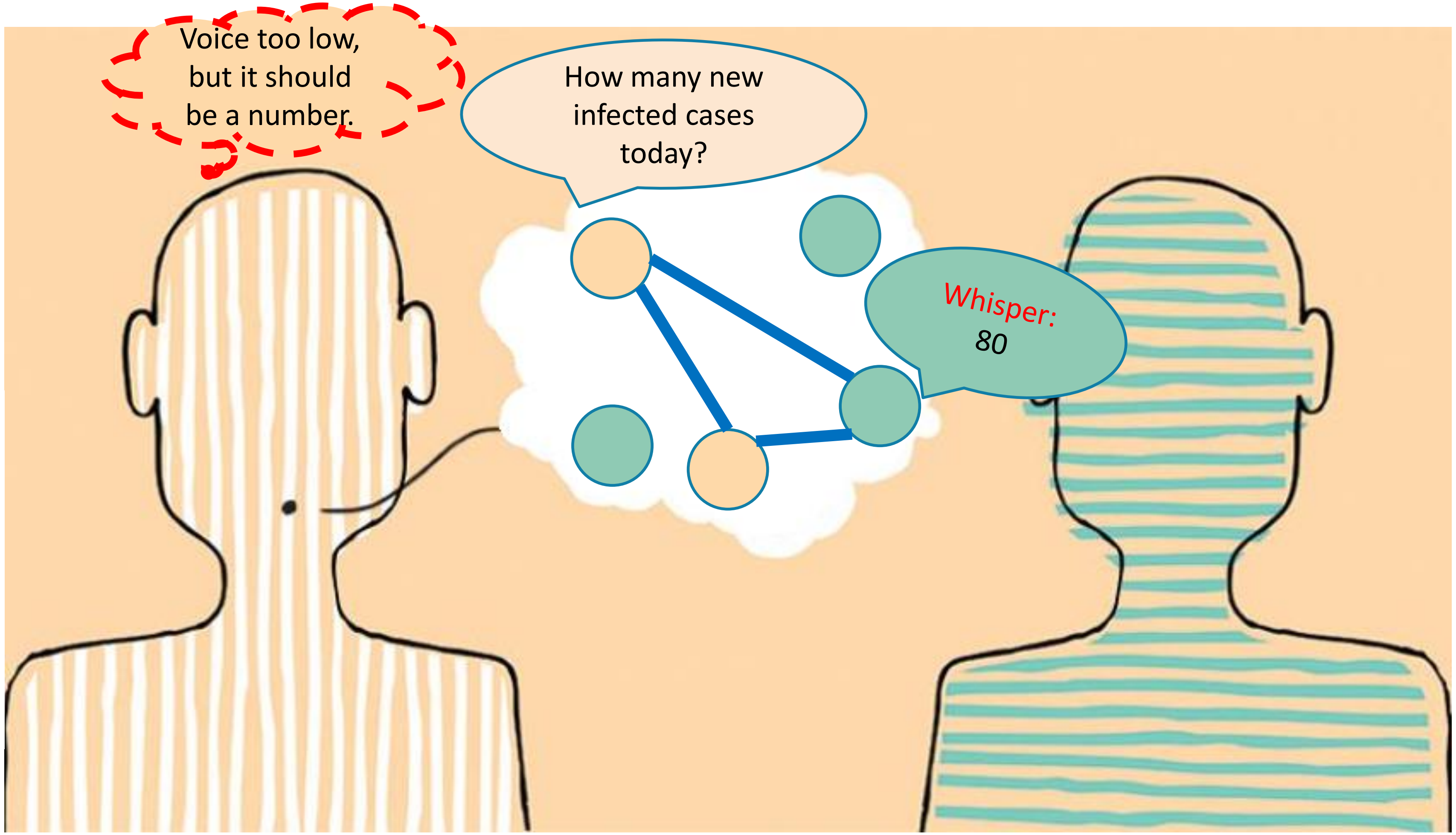
Relational thinking as human speech recognition



Relational thinking as human speech recognition



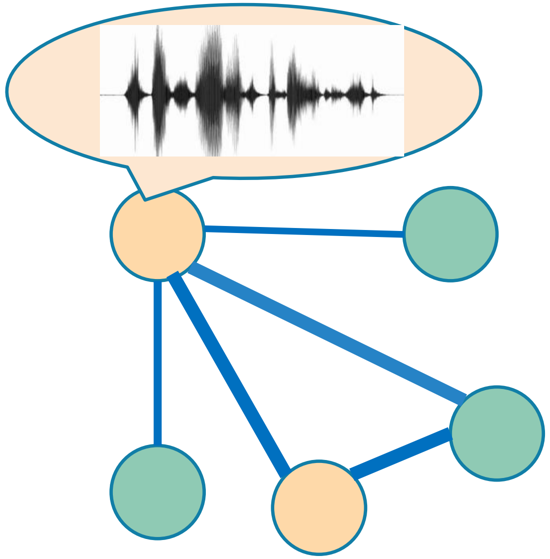
Relational thinking as human speech recognition



Problem formulation

- Given the current utterance $\mathbf{X}_i$ and its histories (of fixed size, for simplicity)
- We aim to simulate relational thinking process, which is embedded into ASR:
  - Construct an infinite number of graphs $\{G^{(k)}\}_{k=1}^{+\infty}$
    - where $G^{(k)}$ represent k-th percept for multiple utterances
  - Then, these percept graphs are combined and further transformed via a graph transform $\mathbf{S}$.
- Our ultimate goal: $P(\mathbf{Y}_i|\mathbf{X}_i, \{G^{(k)}\}_{k=1}^{+\infty}, \mathbf{S})$, with a close form solution
- So that, perception and transformation can be decoupled from speech (graph learning)

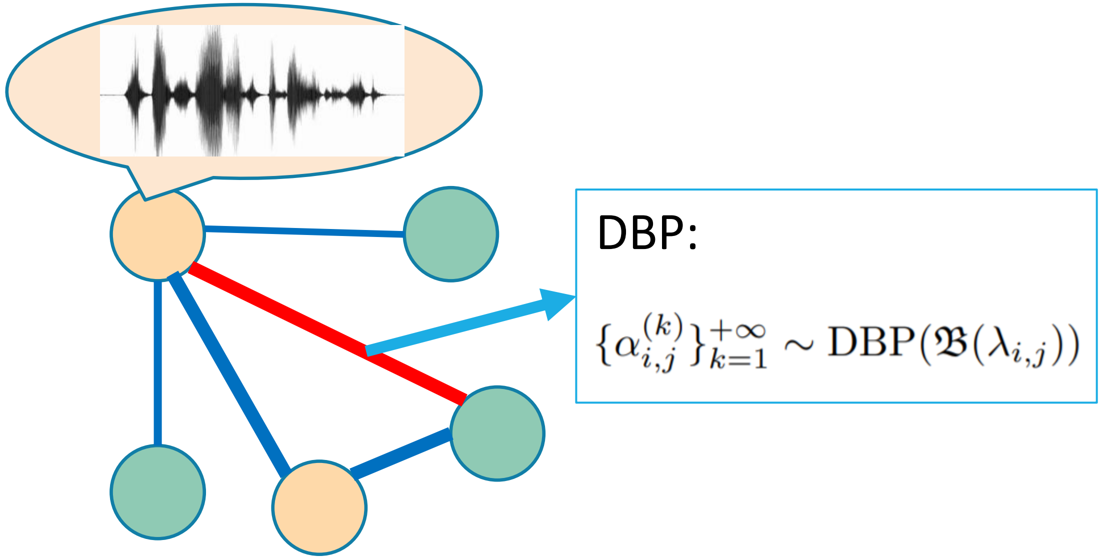
Percept simulator: Deep Graph random process



Deep graph random process (DGP)

- a stochastic process to describe percept generation
- It contains a few nodes, each represents an utterance

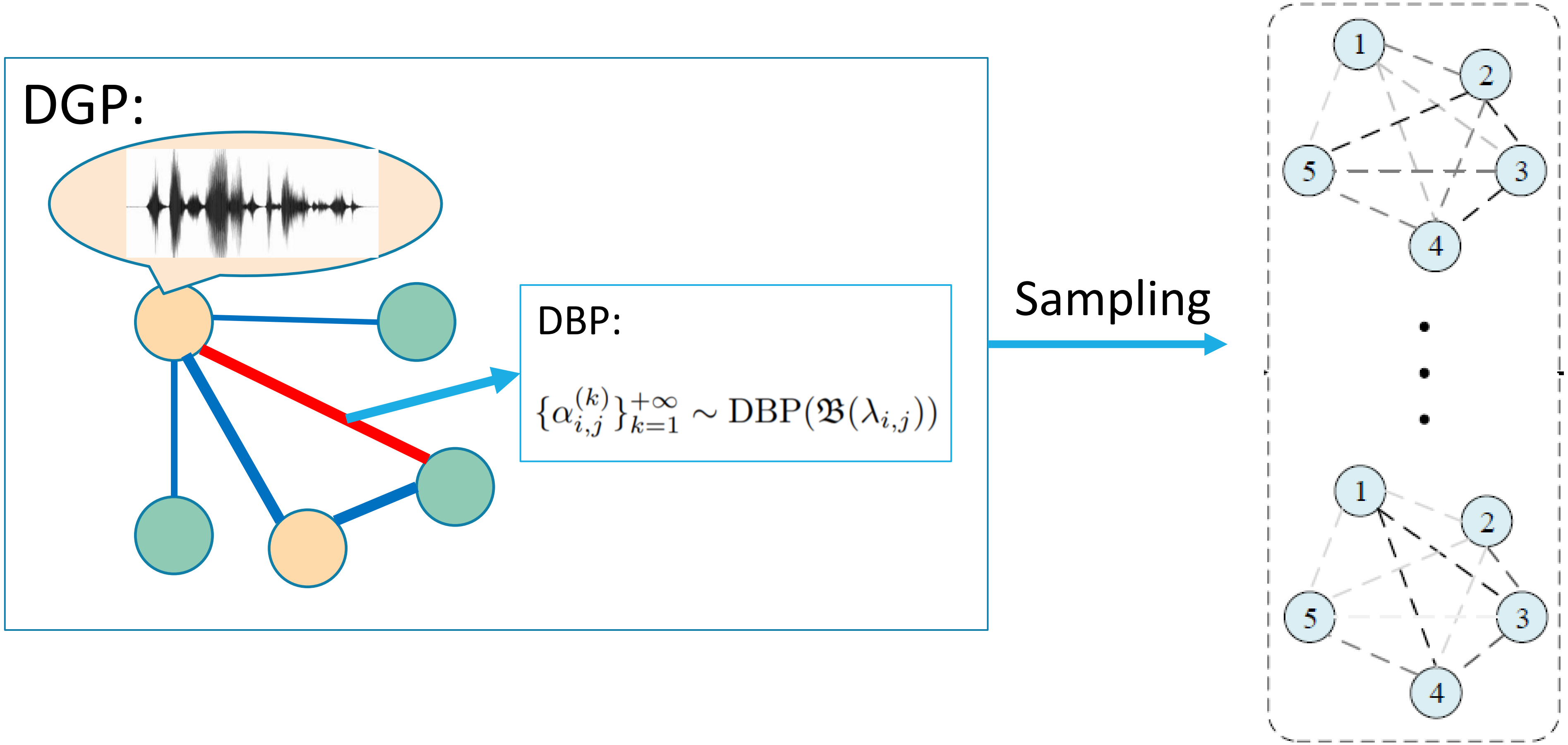
Percept simulator: Deep Graph random process



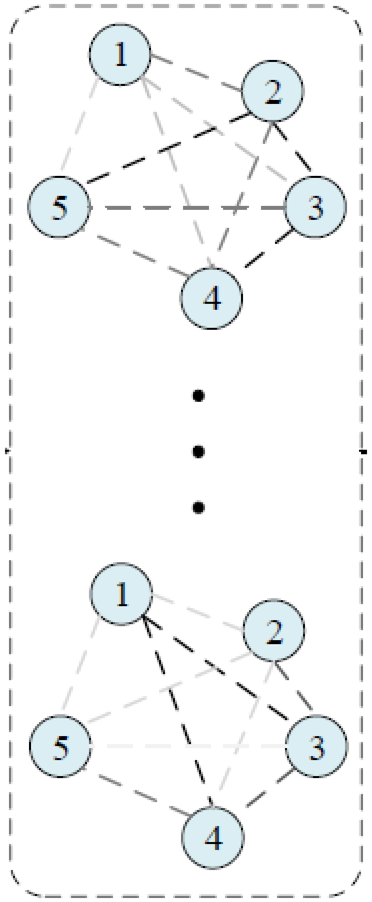
Deep graph random process (DGP)

- a stochastic process to describe percept generation
- It contains a few nodes, each represents an utterance
- Each edge is attached with a deep Bernoulli process (DBP)
  - Special Bernoulli process we proposed
  - Bernoulli parameter $\lambda_{i,j}$ is assumed to be close to 0

Sampling from DGP



Coupling of innumerable percept graphs



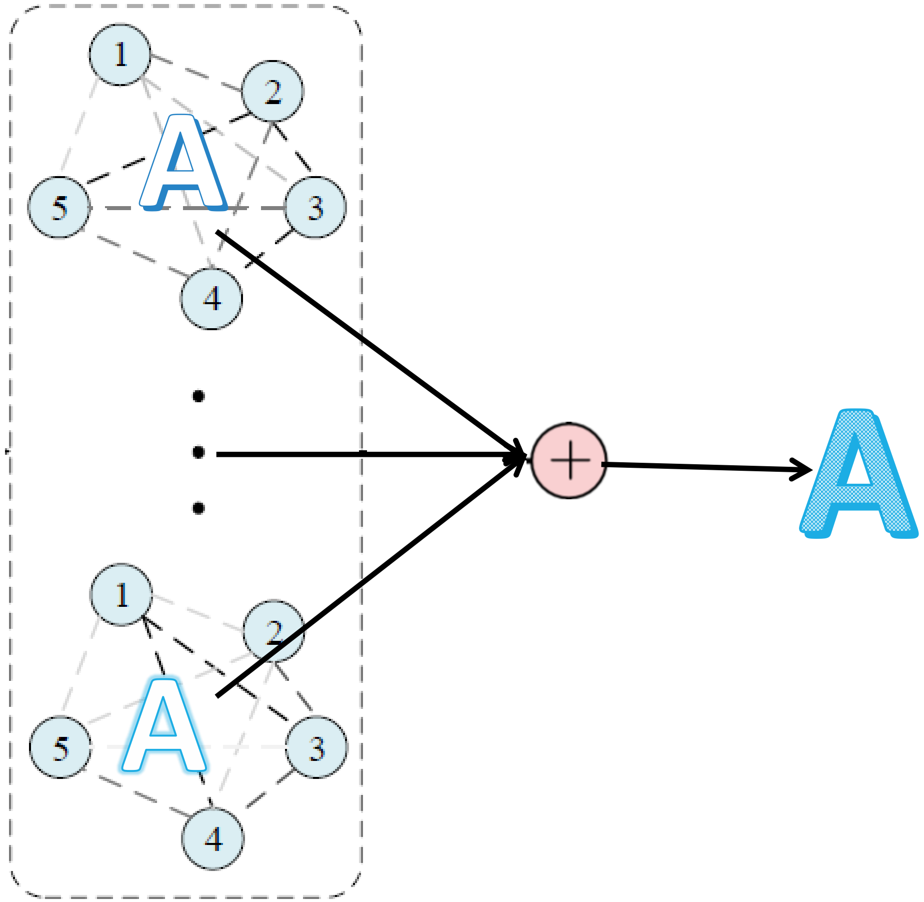
Coupling in DGP

- The goal is to extract a representation of an infinite number of percept graphs

Coupling of innumerable percept graphs

	1	2	3	4	5
1			1	1	
2				1	1
3	1				1
4	1	1			
5		1	1		

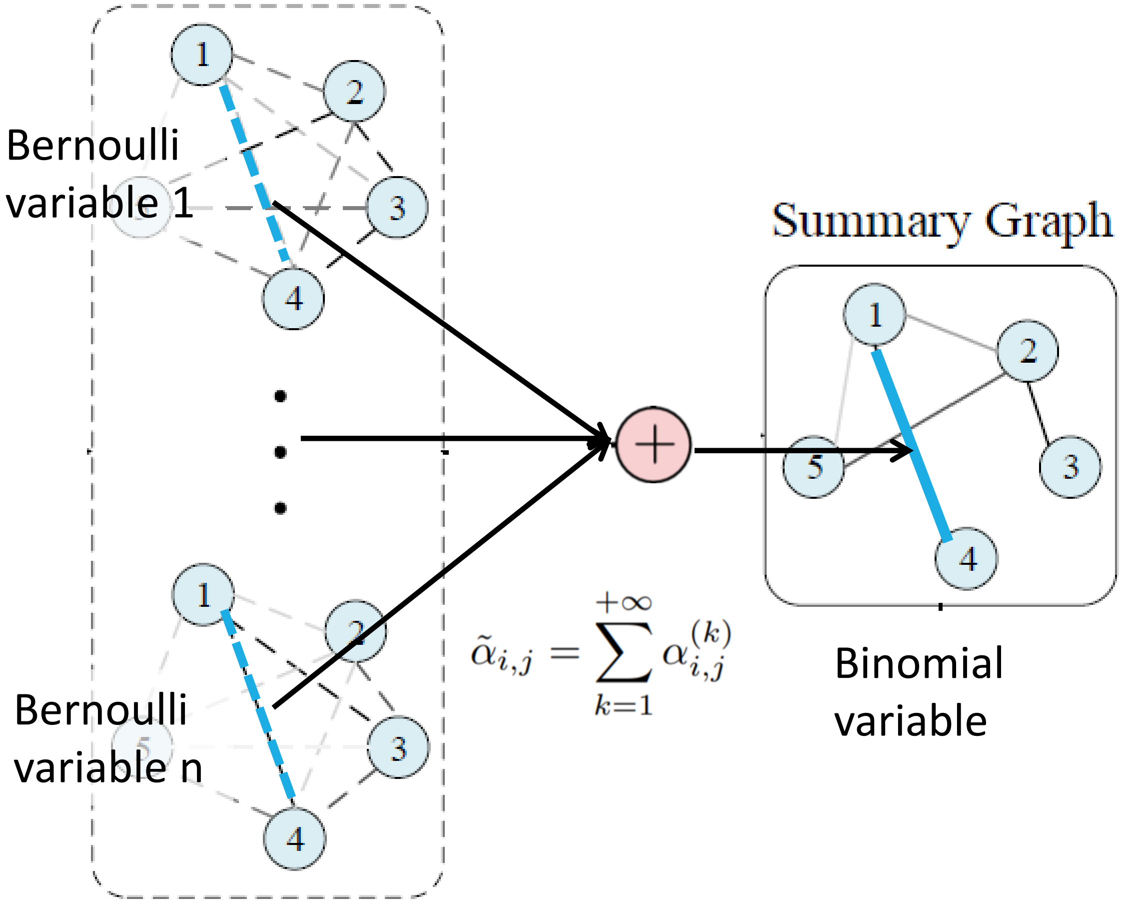
	1	2	3	4	5
1		1			1
2	1		1		
3		1		1	
4			1		1
5	1			1	



Coupling in DGP

- The goal is to extract a representation of an infinite number of percept graphs
- Computationally intractable to summing over their adjacency matrices

Coupling of innumerable percept graphs

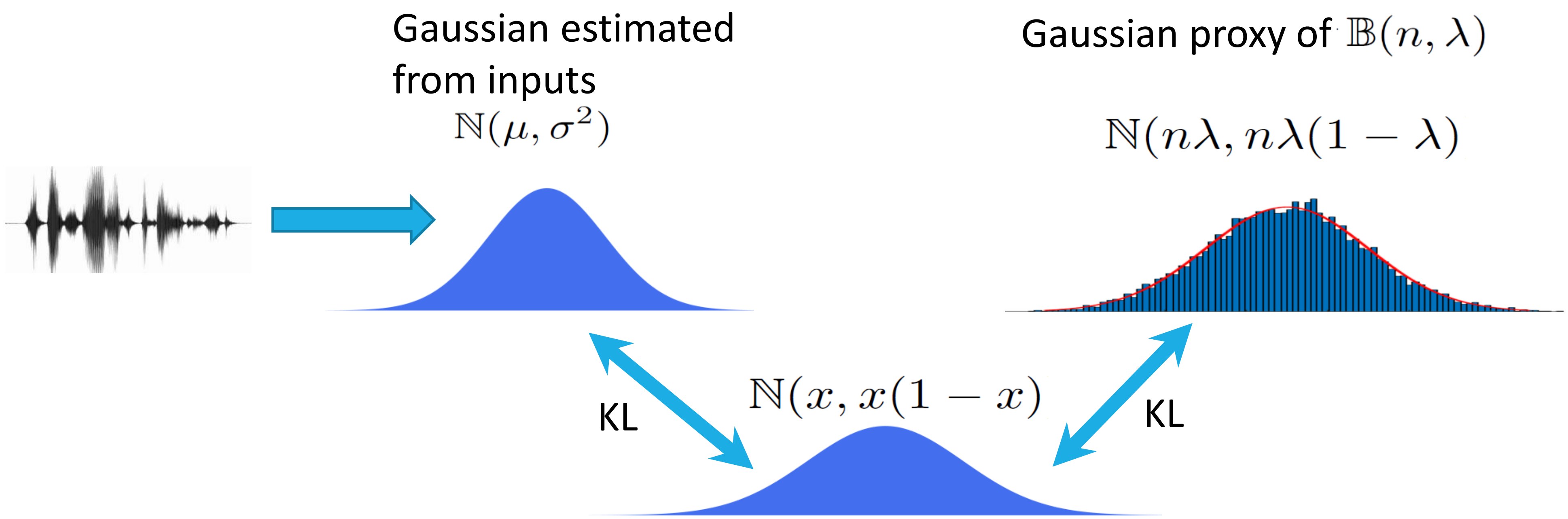


Coupling in DGP

- Construct an equivalent graph
- Summing over the original Bernoulli variables gives a Binomial distribution $\mathbb{B}(n, \lambda)$ with $n \rightarrow +\infty$ and $\lambda \rightarrow 0$.
- Can we inference and sampling from such distribution ?

$\tilde{\alpha}_{i,j} \sim \mathbb{B}(n, \lambda_{i,j})$

Inference and sampling of Binomial distribution with $n \to +\infty$ *and* $\lambda \to 0$



- Approximate above two distributions with bounded appr. errors (Theorem1):

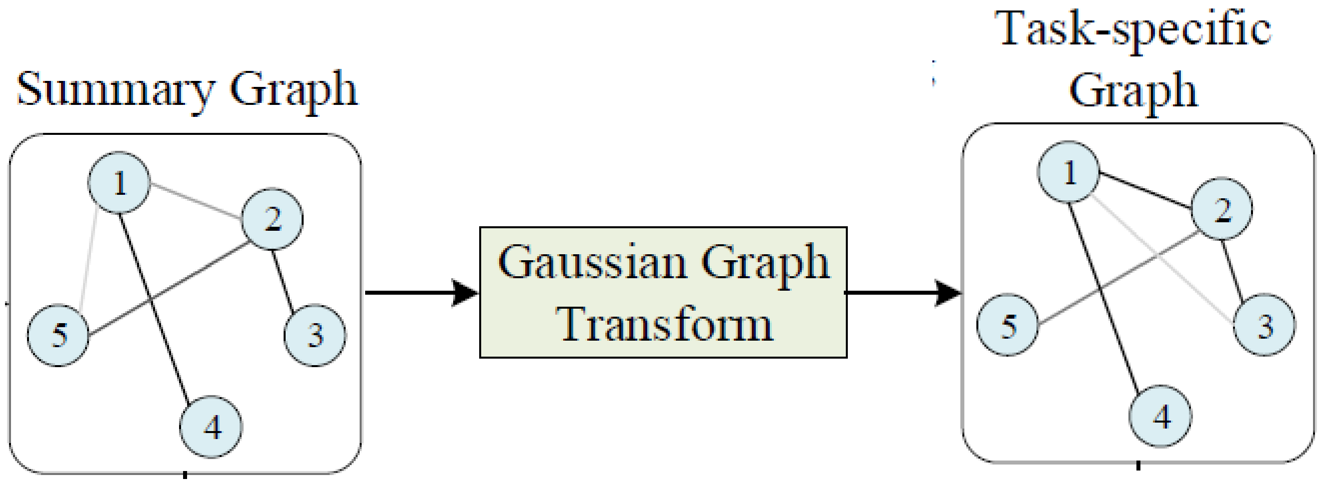
$$x = m = \frac{1 + l - \sqrt{1 + l^2}}{2}, \ where\ l = \frac{2\sigma^2}{1 - 2\mu}$$

Inference and sampling of Binomial distribution with $n \to +\infty$ *and* $\lambda \to 0$

> **Theorem 1 (informal)** Let $\mathbb{B}(n, \lambda)$ denotes an Binomial distribution, with $n \to +\infty$ and $\lambda \to 0$ and let $m = n\lambda$. There exists a Gaussian distribution $\mathbb{N}(m, m(1-m))$ that approximates such Binomial distribution with a bounded approximation error.

- Directly parameterization of n and λ are avoided
- Sampling: this allows for the re-parametrization trick to be used

Transforming the general summary graph to be task-specific

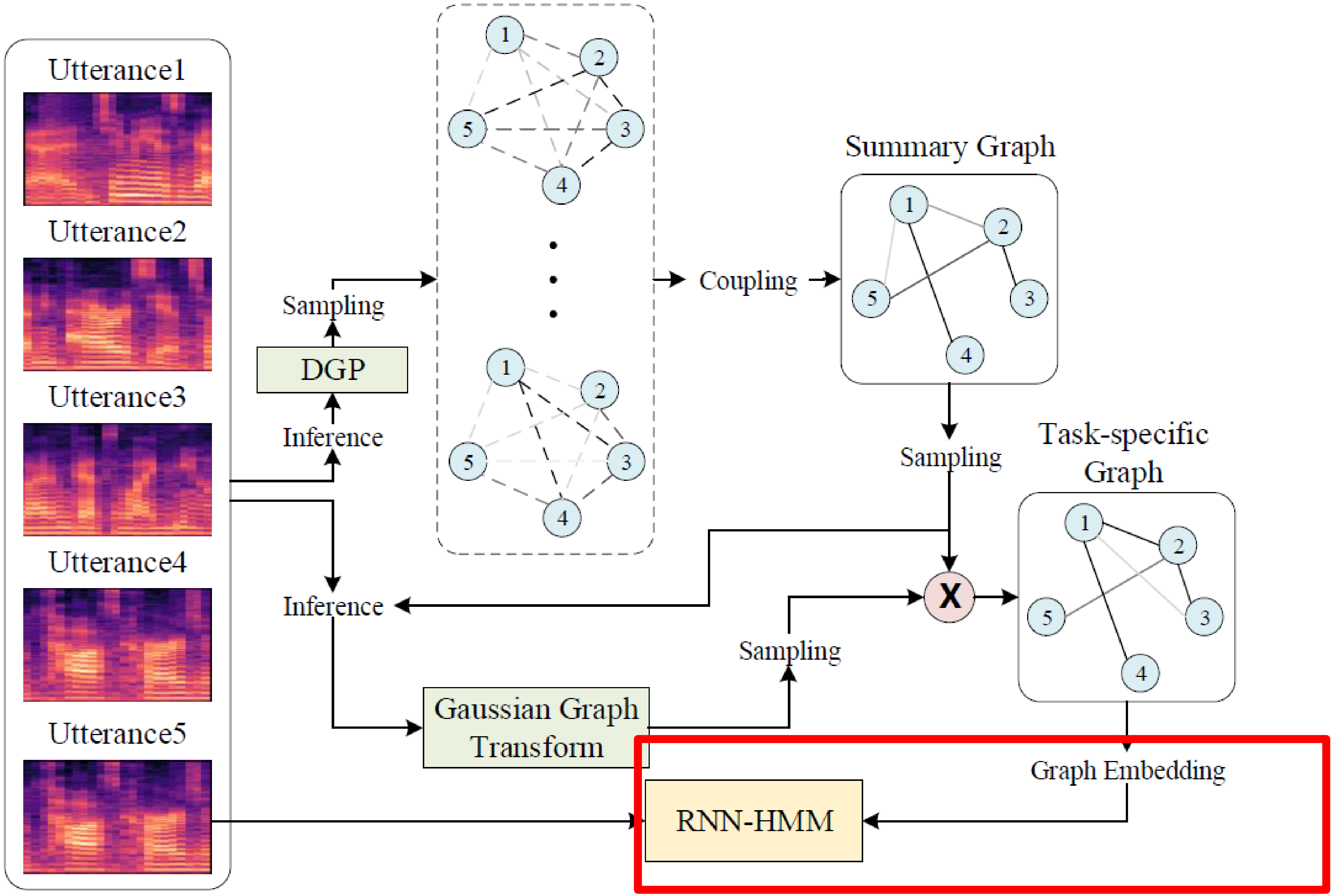


Gaussian graph transform

- Each entry of its transform matrix follows a conditional Gaussian distribution
- Conditioning on edges of summary graph

Application of DGP for acoustic modeling

Relational thinking network (RTN)



Learning

Variational inference is applied to jointly optimise DGP, the Gaussian graph transform, and the RNN-HMM acoustic model

- **Challenge #1** : DGP contains too many latent variables
  - Bernoullis and Binomials are equivalent , specifying one determine the whole DGP

Learning

Variational inference is applied to jointly optimise DGP, the Gaussian graph transform, and the RNN-HMM acoustic model

- **Challenge #1** : DGP contains too many latent variables
  - Bernoullis and Binomials are equivalent , specifying one determine the whole DGP
- **Challenge #2** : One of a KL term of our ELBO is computational intractable

$$\sum_{(i,j)\in\tilde{E}} \boxed{\{\mathrm{KL}(\mathbb{B}(n,\tilde{\lambda}_{i,j})||\mathbb{B}(n,\tilde{\lambda}_{i,j}^{(0)})} \rightarrow \text{This is computational intractable, as n approaches infinity}$$

The analytical evidence lower bound (ELBO)

Theorem 2 (informal) Suppose we are given a DGP consisting of a summary graph whose edge follows Binomial distribution $\mathbb{B}(n, \lambda_{i,j})$ with $n \to +\infty$ and $\lambda_{i,j} \to 0$. There exists a close form solution for ELBO of DGP, which is irrelevant to the infinity n.

- This theorem allows us to obtain a close form solution of ELBO.
- In particular:

$$\mathrm{KL}(\mathbb{B}(n, \tilde{\lambda}_{i,j}) || \mathbb{B}(n, \tilde{\lambda}_{i,j}^{(0)}) < m_{i,j} \log \frac{m_{i,j}}{m_{i,j}^{(0)}} + (1 - m_{i,j}) \log \frac{1 - m_{i,j} + m_{i,j}^2/2}{1 - m_{i,j}^{(0)} + {m_{i,j}^{(0)}}^2/2}$$

- The solution is irrelevant to the infinity n

Experiments: data sets

We evaluated the proposed method on several ASR datasets:

ASR tasks

- CHiME-2 (preliminary study, not a conversational ASR task):
  - Noisy version of WSJ0
- CHiME-5 (conversaitional ASR task)
  - First large-scale corpus of real multi-speaker conversational speech
  - Train: ~40 hours, Eval: ~5 hours.

Quantitative/qualitative study of the generated graphs

- Synthetic Relational SWB
  - SWB: telephony conversational speech
  - SwDA: extends SWB with graph annotations for utterances
  - Train: 30K utterances (without graphs) , Test: graphs involved in 110K utterances

Experiments: model configurations

L: number of layers;
N: number of hidden states per layer;
P: number of model parameters
T: training time per epoch (hrs)

Model	L	N	P	T
LSTM (Huang et al., 2019)	3	2048	130M	0.71
SRU (Huang et al., 2019)	12	2048	156M	0.32
RPPU (Huang et al., 2019)	12	1024	142M	0.37
Our SRU (Lei et al., 2017)	12	1280	63M	0.09
VSRU (Chung et al., 2015)	9	1024	66M	0.09
RRN (Palm et al., 2018)	9	1024	64M	0.09
RTN (Ours)	9	1024	70M	0.11

Hengguan Huang, Hao Wang, Brain Mak. Recurrent Poisson process unit for speech recognition. AAAI, 2019.

Robustness to input noise

Detailed WER (%) on test set of CHiME-2

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

ASR Results on conversational task

WER (%) Eval of CHiME5

Model	WER
Kaldi DNN (Povey et al., 2011b)	64.5
SRU (Lei et al., 2017)	62.6
VSRU (Chung et al., 2015)	61.6
RTN (Ours)	**57.4**

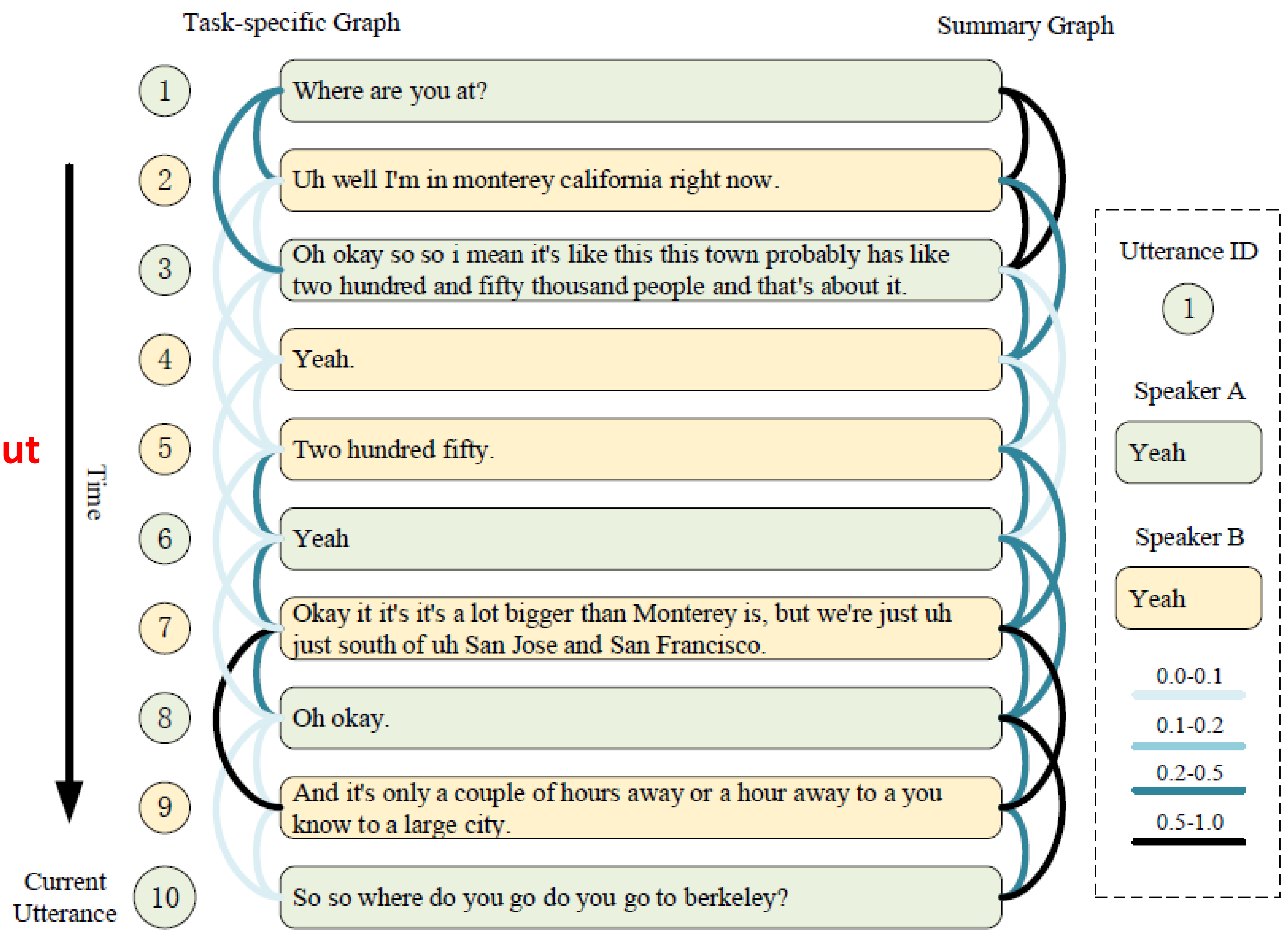
Outperforms other baselines

Quantitative study: can we infer utterance relationships with the generated graphs

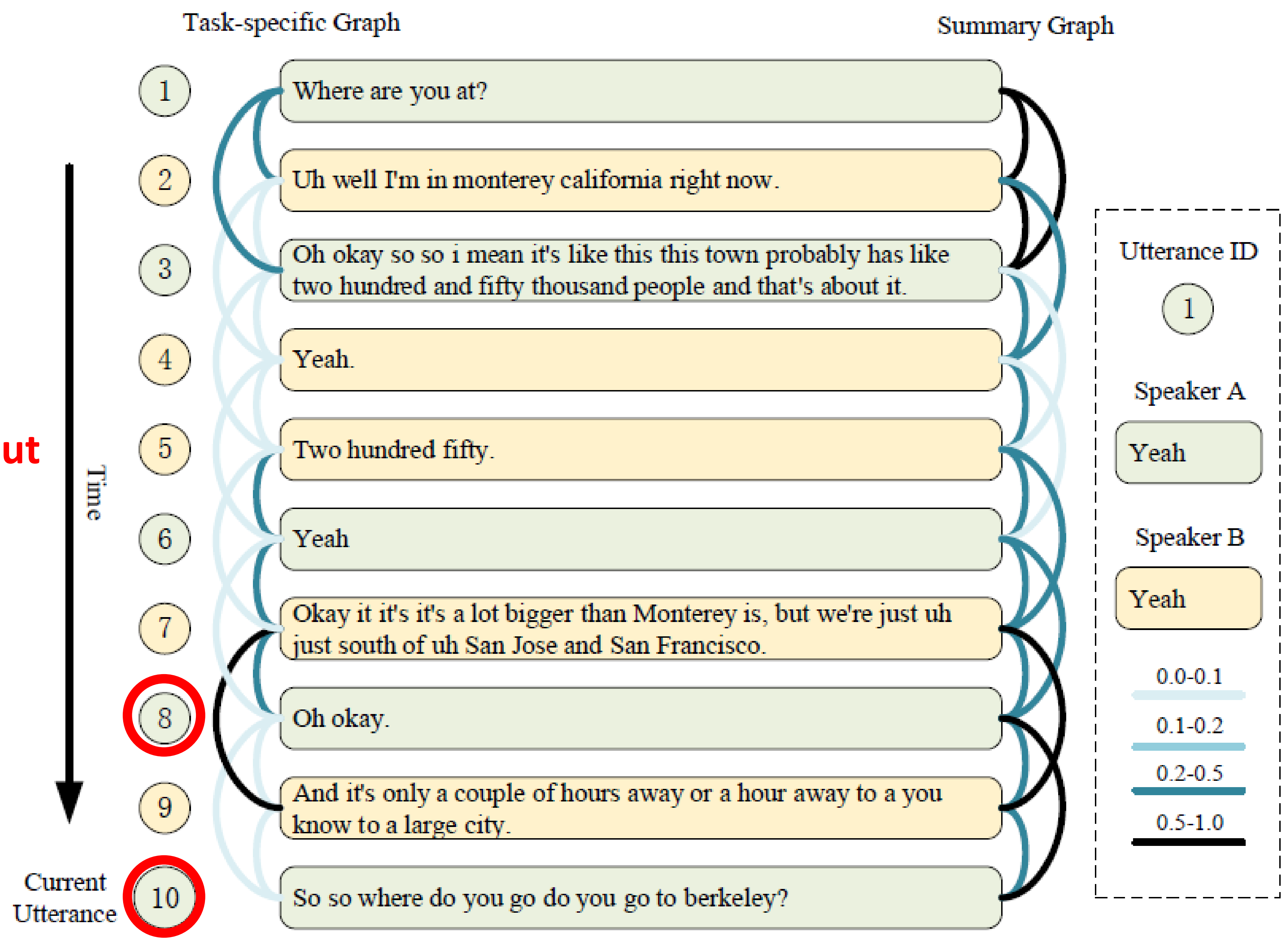
Error rate(%) of relation prediction on Synthetic Relational SWB

Graph Type	Err
Random Graph	50.0
Summary Graph	28.6
Task-specific Graph	28.7

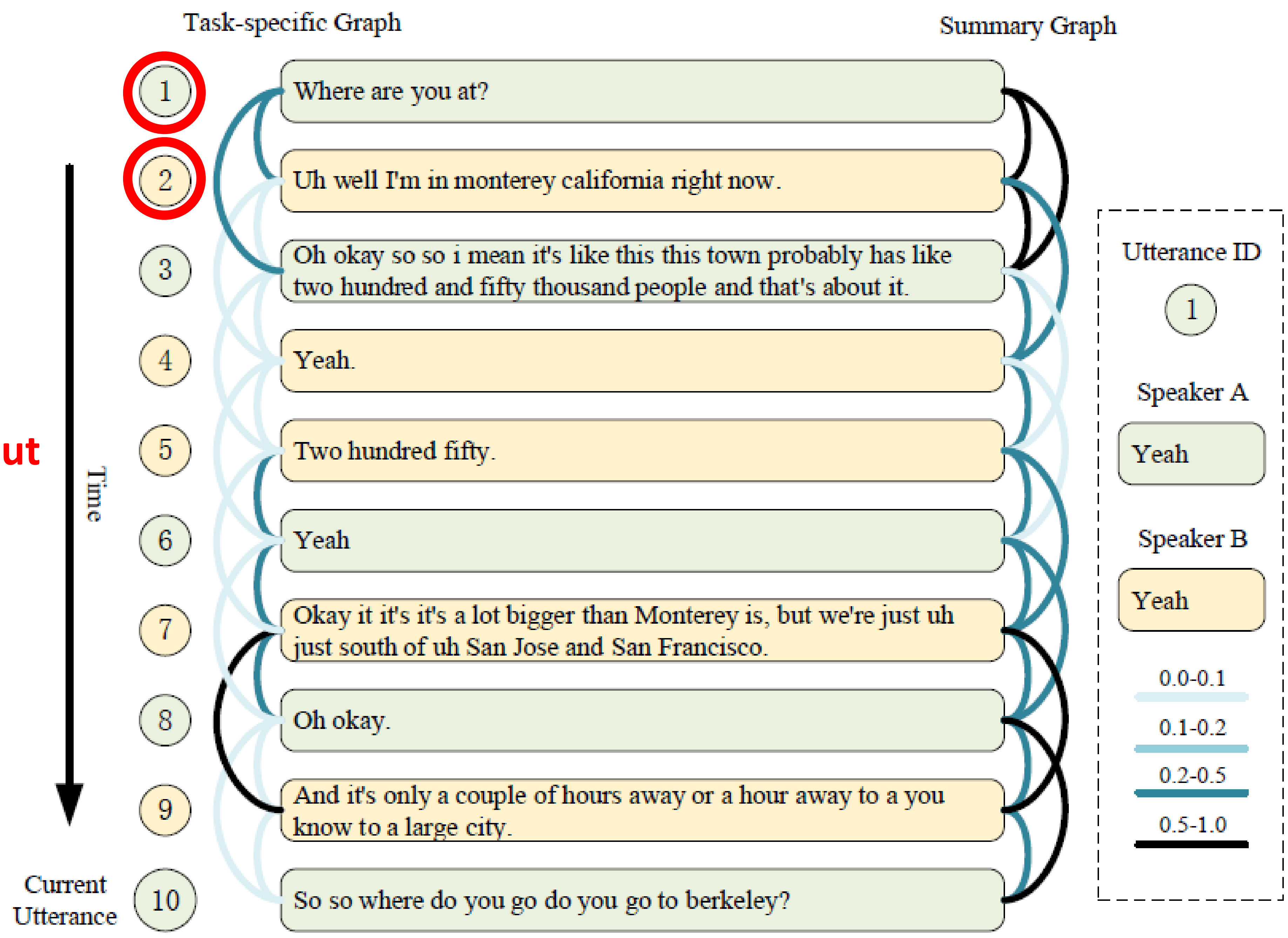
We can capture relationships without relational data !



We can capture relationships without relational data !



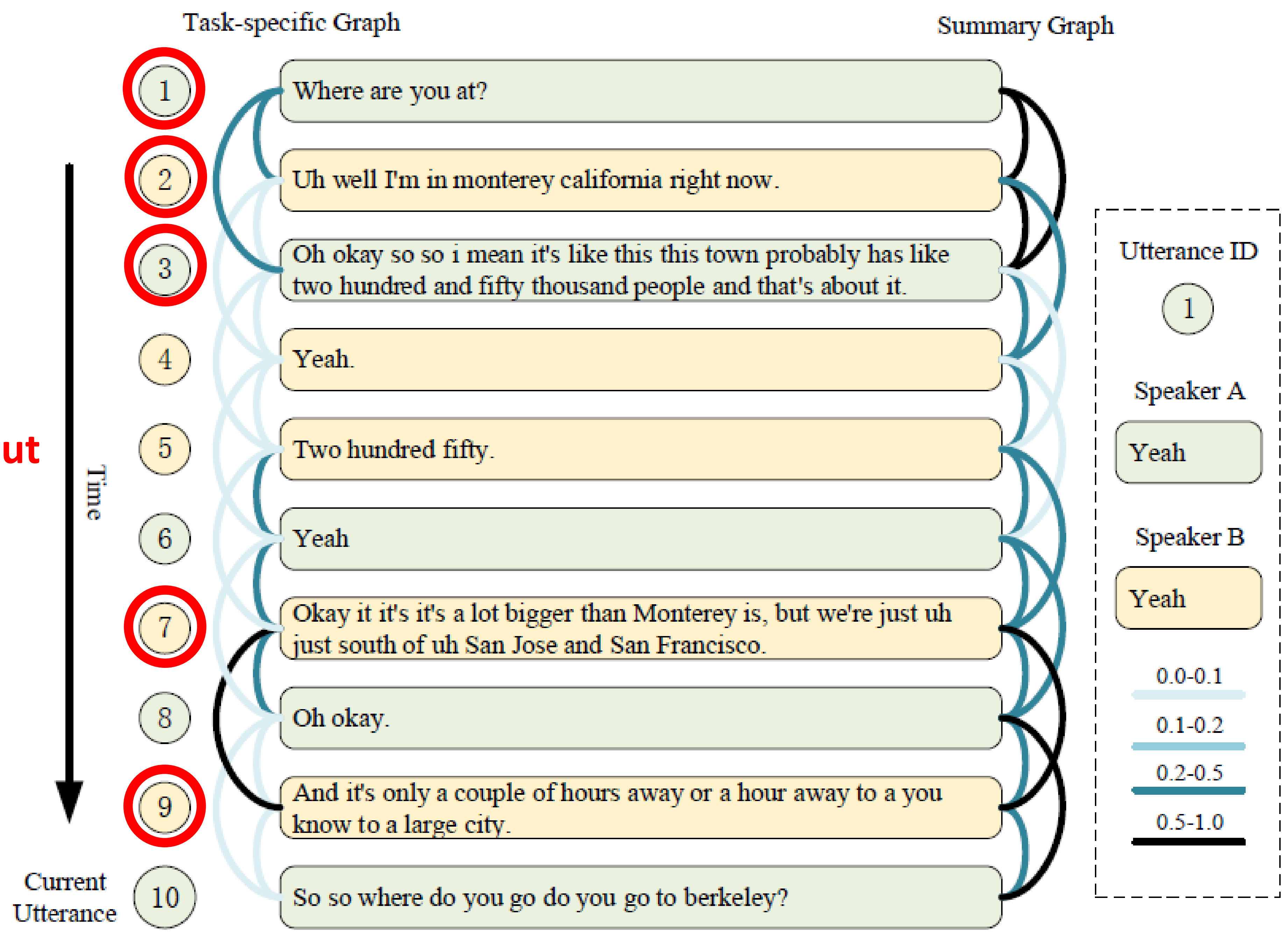
We can capture relationships without relational data !



Recognition results of the utterance 10

Ground truth: so so where do you go do you go to Berkeley
SRU: so so what do you go do you go to Berkeley
RTN (ours): so so where do you go do you go to Berkeley

We can capture relationships without relational data !



Take-away

Expand the variational family with a deep graph random process

- Enable relational thinking modelling
- Graph learning without any relational labelling
- Easy to be applied for a downstream task such as ASR
- Improvements on several speech recognition datasets
- Code (coming soon): https://github.com/GlenHGHUANG/Deep_graph_random_process