2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU2021)

Low-Latency Incremental Text-to-Speech Synthesis with Distilled Context Prediction Network

Takaaki Saeki, Shinnosuke Takamichi, Hiroshi Saruwatari

The University of Tokyo, Japan

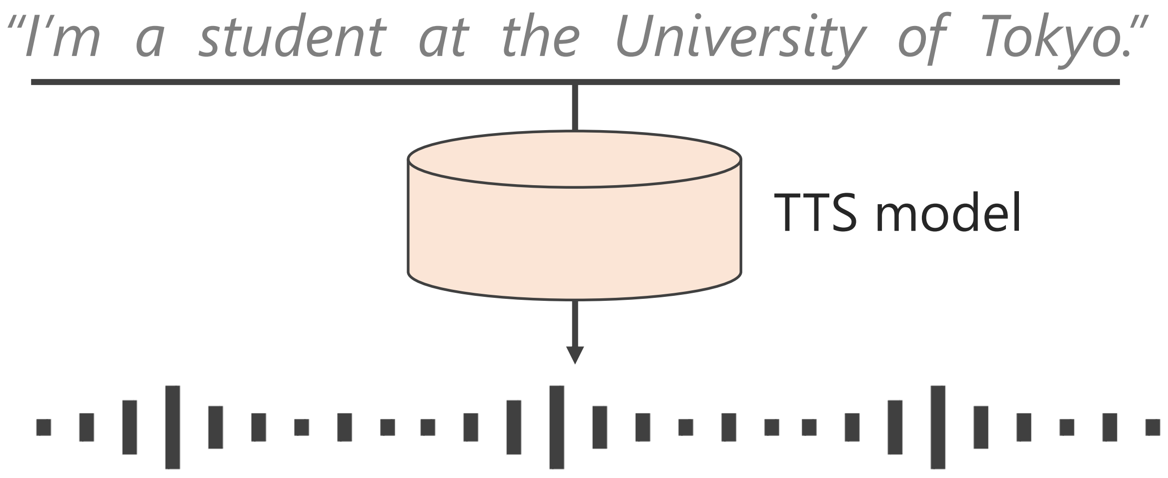
MOONSHOT RESEARCH & DEVELOPMENT PROGRAM

AVATAR SYMBIOTIC SOCIETY

Incremental Text-to-Speech Synthesis

High-quality and low-latency streaming TTS is needed for real-time speech generation

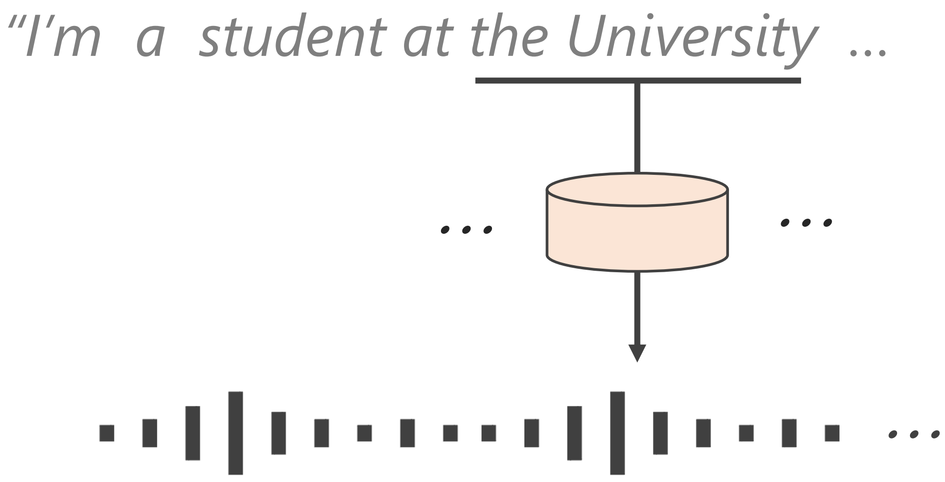
Sentence-level TTS



Synthetic speech

High naturalness but requiring **latency** due to full-sentence observation

Incremental TTS



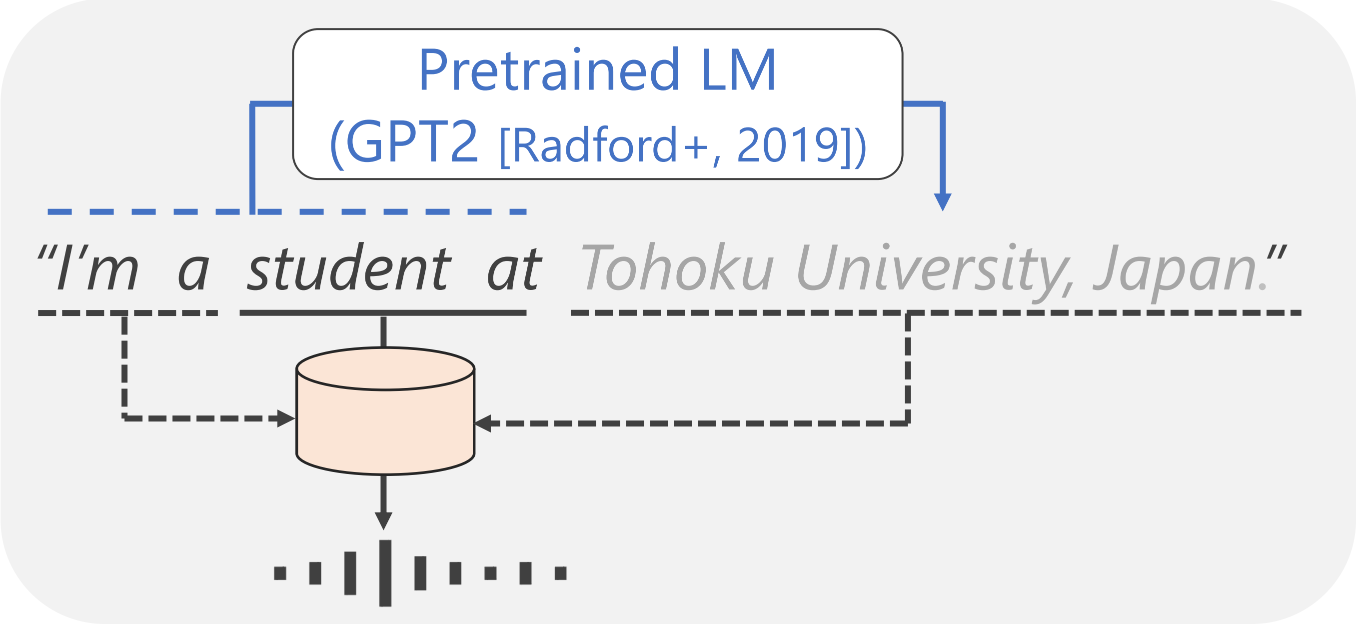
Synthesizing **without using full sentence**

Tradeoff between quality and latency

Incremental TTS with Pseudo Lookahead

Related work 1 [Saeki+, 2021]

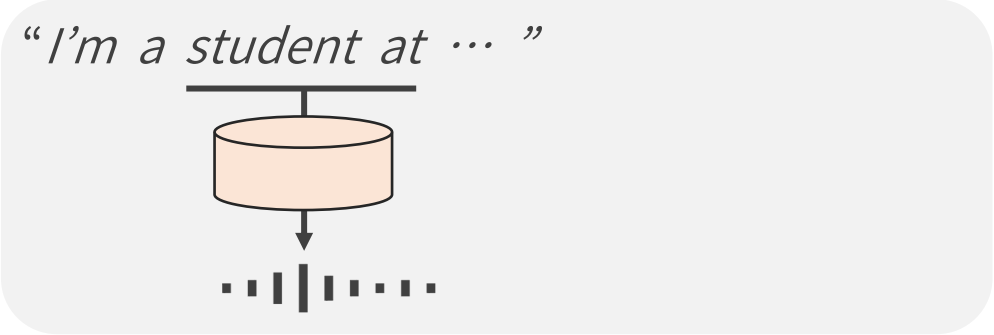
Pseudo lookahead with language model
Synthesizing current speech segment with **unobserved future context**
(Imitating human's incremental reading)



Achieving high naturalness without waiting for observation of future context

Related work 2 [Yanagita+, 2019]

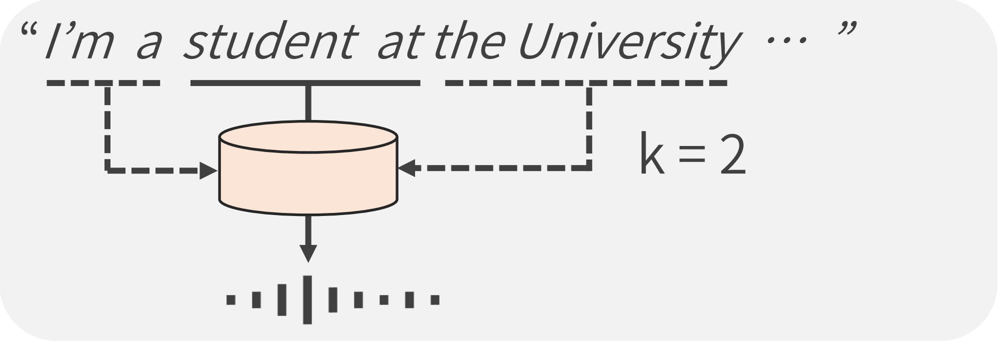
Generating each segment independently



Low-quality output speech

Related work 3 [Ma+, 2020]

Waiting for k words (lookahead-*k* policy)



Need waiting time of subsequent words

Motivation and Concept of Proposed Method

Challenge of previous method using pseudo lookahead with language model

- **Huge processing time** due to inference of GPT-2 at each time step
- Need to achieve **faster speech synthesis than human's speaking speed**

Proposed method: **Fast listen-while-predict framework with language model distillation**

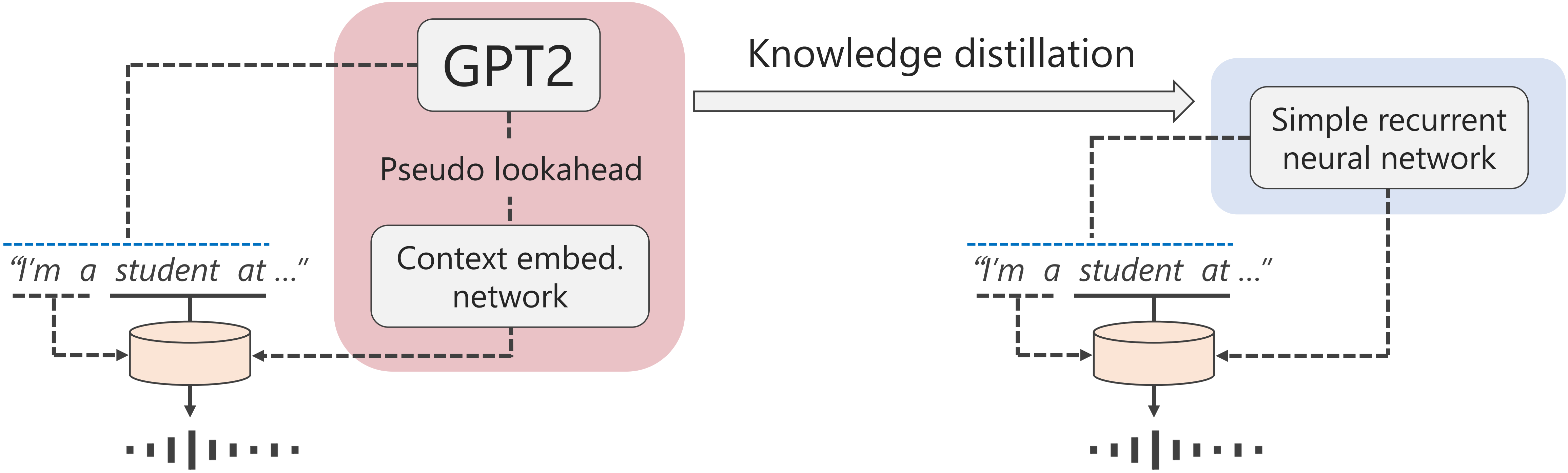
- knowledge distillation from GPT2 + contextual embedding to single lightweight model
- **Directly predicting future context** from observed words for **fast inference**

Overview of Proposed Method

Inspired by **task-specific knowledge distillation of language model** (BERT) [Tang+, 2019]

- Distilling from BERT to lightweight recurrent model without attention
- Student model achieves **comparable performance** to Teacher model in various tasks

Our proposed method aims to perform task-specific knowledge distillation of GPT2 for context estimation task of text-to-speech synthesis



Proposed Teacher-Student Training Framework

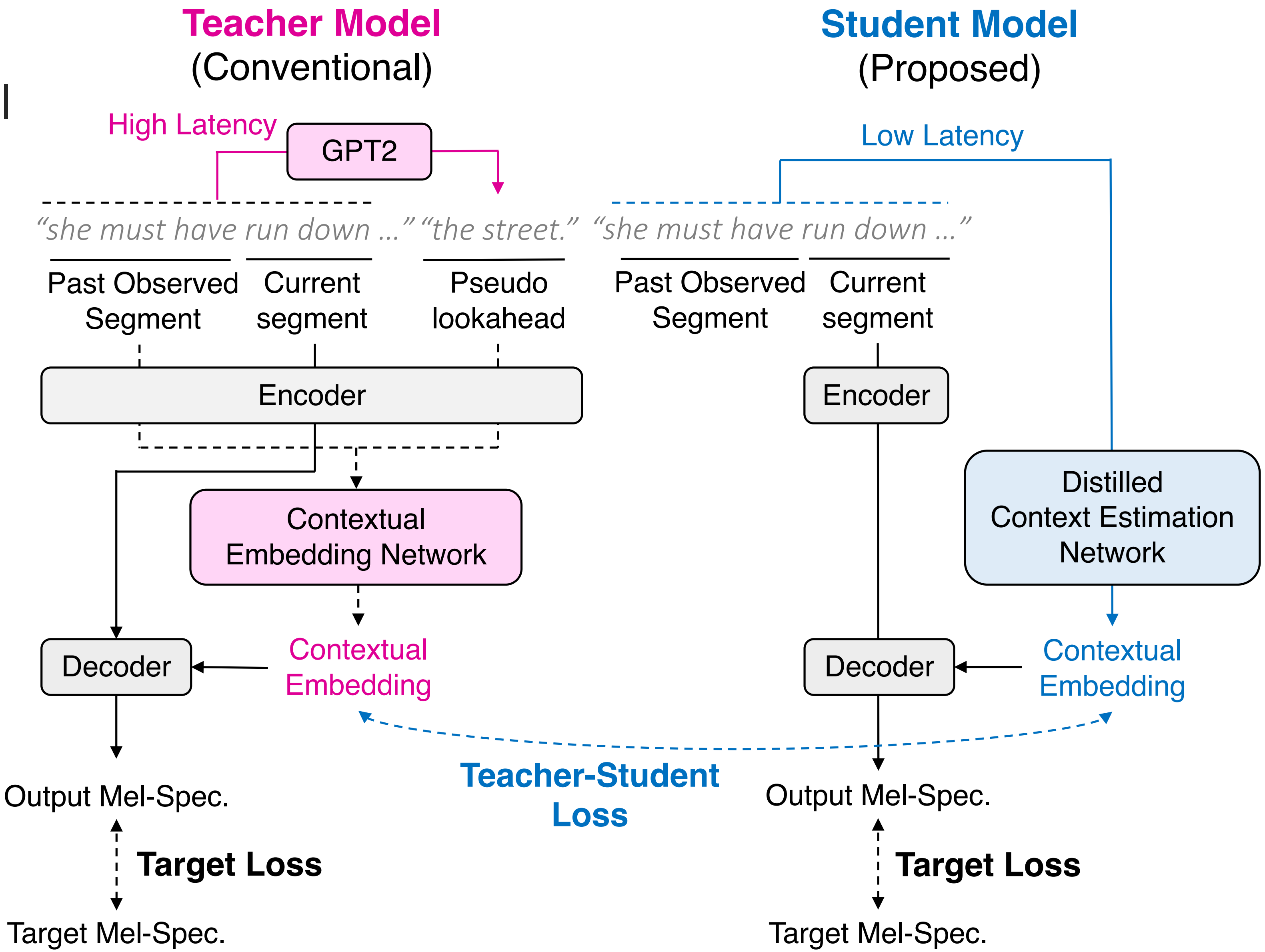
Distilling from previous Teacher model to lightweight Student model

Predicting contextual embedding with single distilled context estimation network

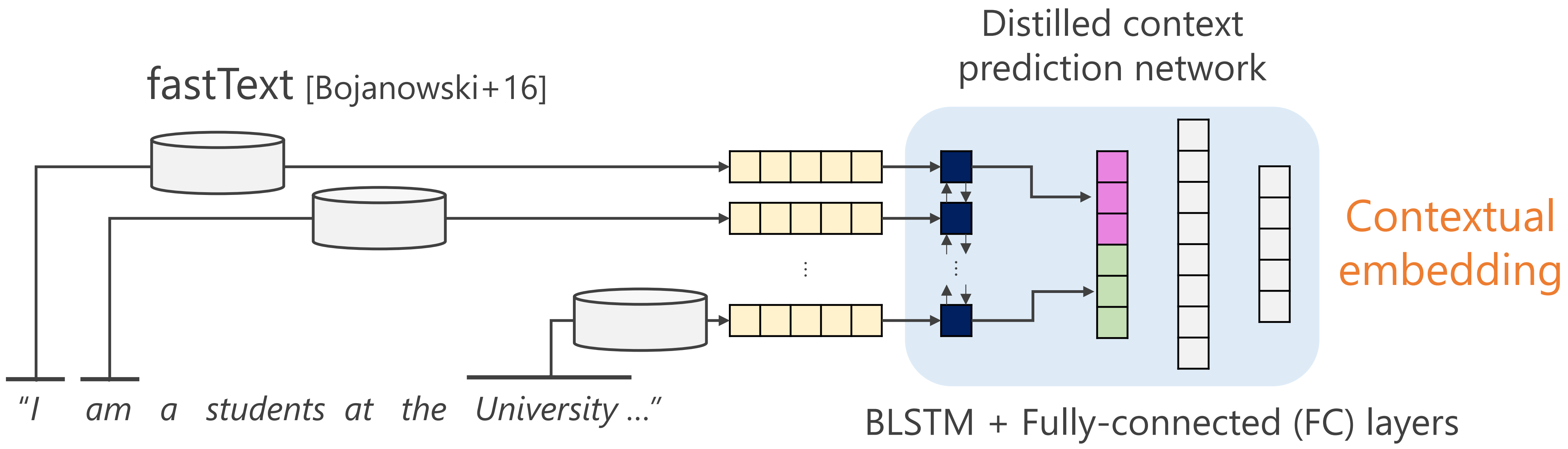
Teacher-Student loss between contextual embedding vectors

Defining objective function with Teacher-Student loss and target loss

$$\mathcal{L} = (1-\lambda)\cdot\mathcal{L}_{\text{target}} + \lambda\cdot\mathcal{L}_{\text{distil}}$$



Architecture of Distilled Context Prediction Network



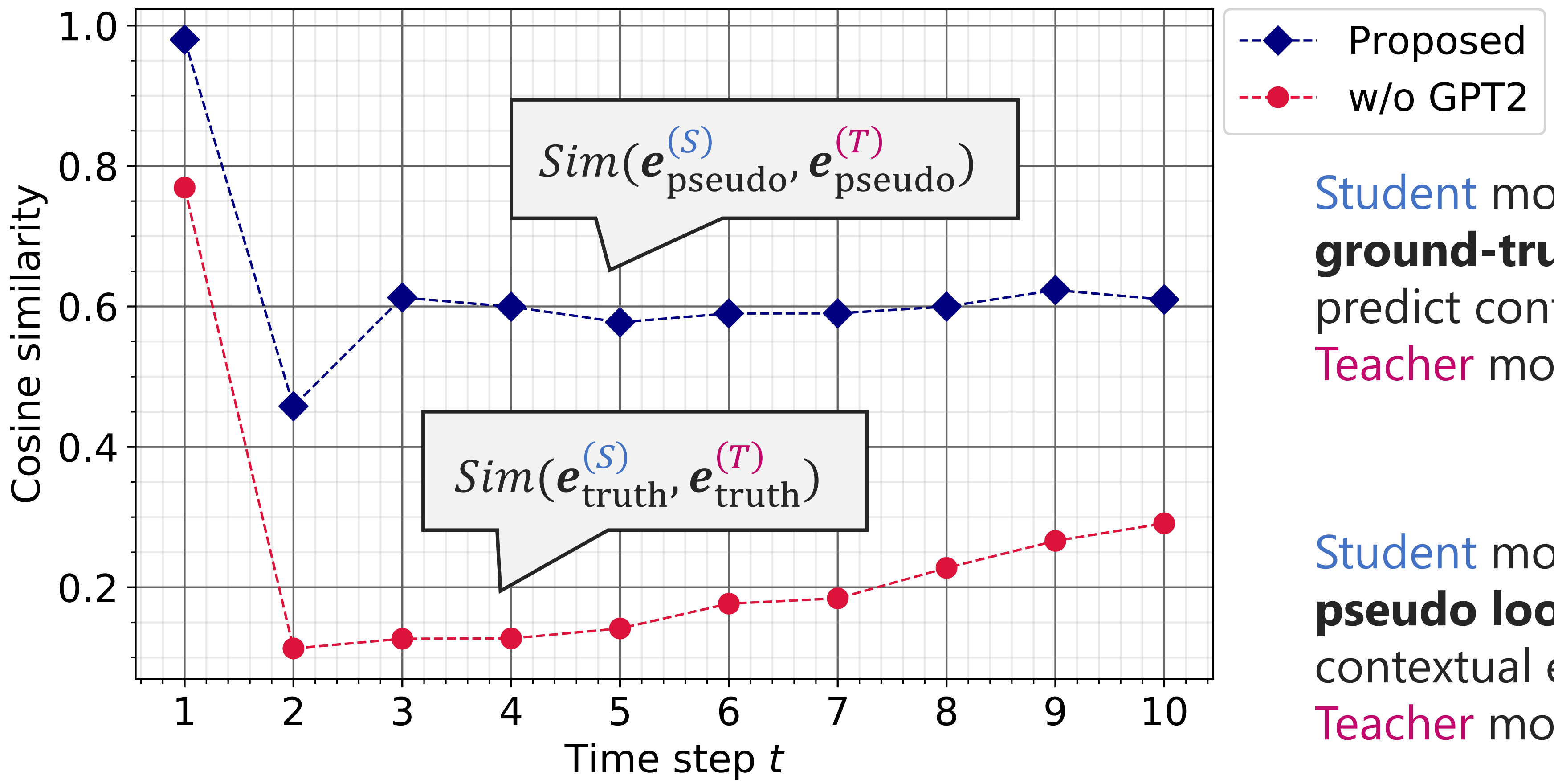
Obtaining distributed representation of observed words with pretrained fastText [Bojanowski+, 2016]

Estimating contextual embedding with lightweight recurrent model without attention

Compared three model sizes: *small*, *medium*, *large*

- (BLSTM-hidden, FC-hidden) = {(100, 200), (300, 600), (500, 1000)}

Discussion on Knowledge Distillation with Language Model



Student model obtained with **ground-truth lookahead** cannot predict contextual embedding of Teacher model

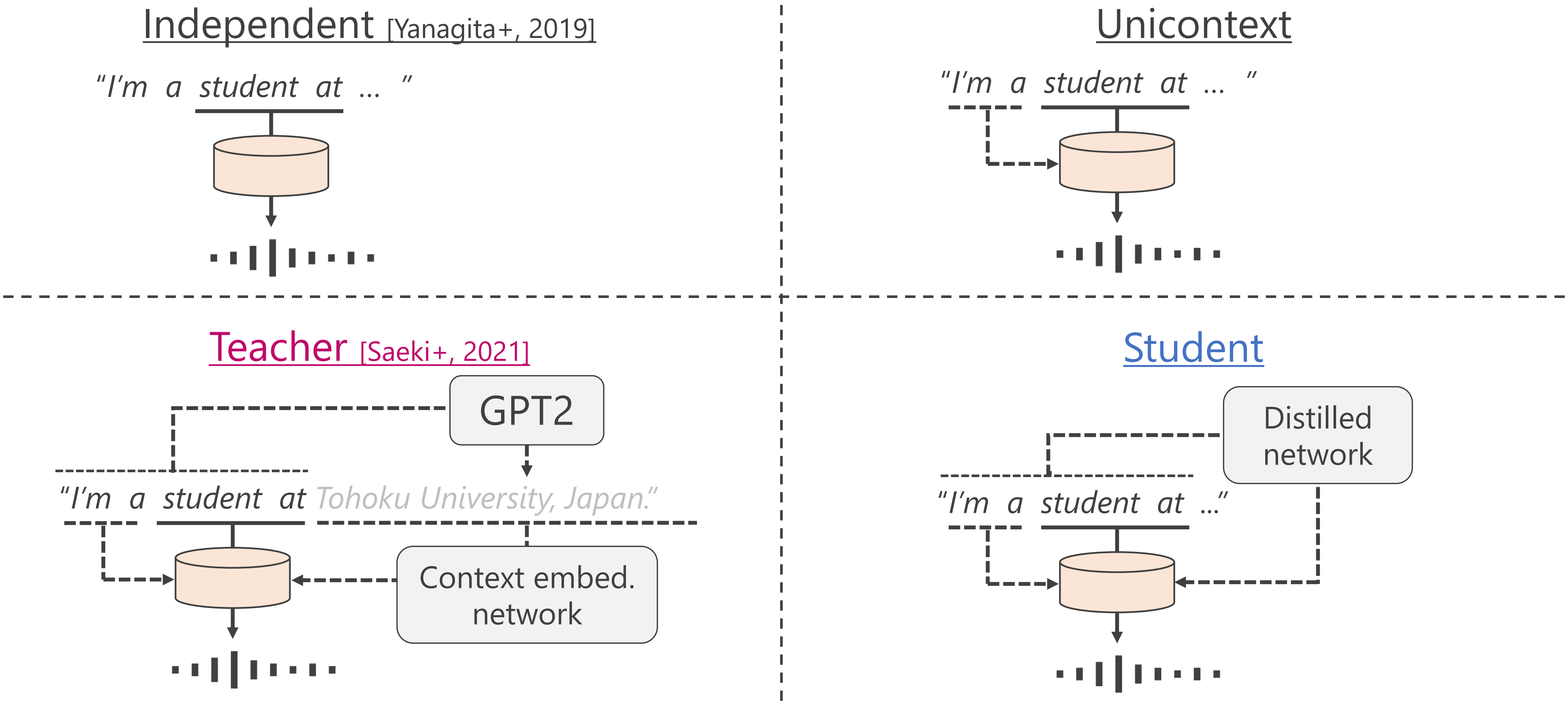
⇕

Student model obtained with **pseudo lookahead** can predict contextual embedding of Teacher model with higher similarity

$\boldsymbol{e}_{\text{pseudo}}^{(S)}$: Contextual embedding with Student model trained using **pseudo lookahead**
$\boldsymbol{e}_{\text{pseudo}}^{(T)}$: Contextual embedding with Teacher model trained using **pseudo lookahead**
$\boldsymbol{e}_{\text{truth}}^{(S)}$: Contextual embedding with Student model trained using **ground-truth lookahead**
$\boldsymbol{e}_{\text{truth}}^{(T)}$: Contextual embedding with Teacher model trained using **ground-truth lookahead**

Experimental Evaluation

Corpus: LJSpeech [Ito+, 2017] (22.05 kHz)



Evaluation Results on Synthetic Speech quality

Objective evaluation: Calculated character error rate (CER) and word error rate (WER) of synthetic speech

Subjective evaluation: Mean opinion score (MOS) test on naturalness evaluated by 40 native speakers

	Full sentence	Unicontext	Teacher	Student w/o target loss	Student w/ target loss
CER (↓)	5.5 %	20.8 %	7.8 %	8.4 %	12.7 %
WER (↓)	18.2 %	49.4 %	22.2 %	22.2 %	33.8 %
MOS (↑)	3.82	3.10	3.51	3.47	3.39

Student > Unicontext & Student ≈ Target: Student model predicted effective contextual embedding for incremental TTS and **achieved comparable naturalness** to Teacher model

Student model performed better without target loss (correspond to results in previous work [Tang+, 2019])

Evaluation Results on Inference Speed

Independent, Unicontext ≈ 0.15s / step
Teacher ≈ 1.5s / step
Student ≈ 0.15s / step

Student achieved around 10 times faster inference than Teacher

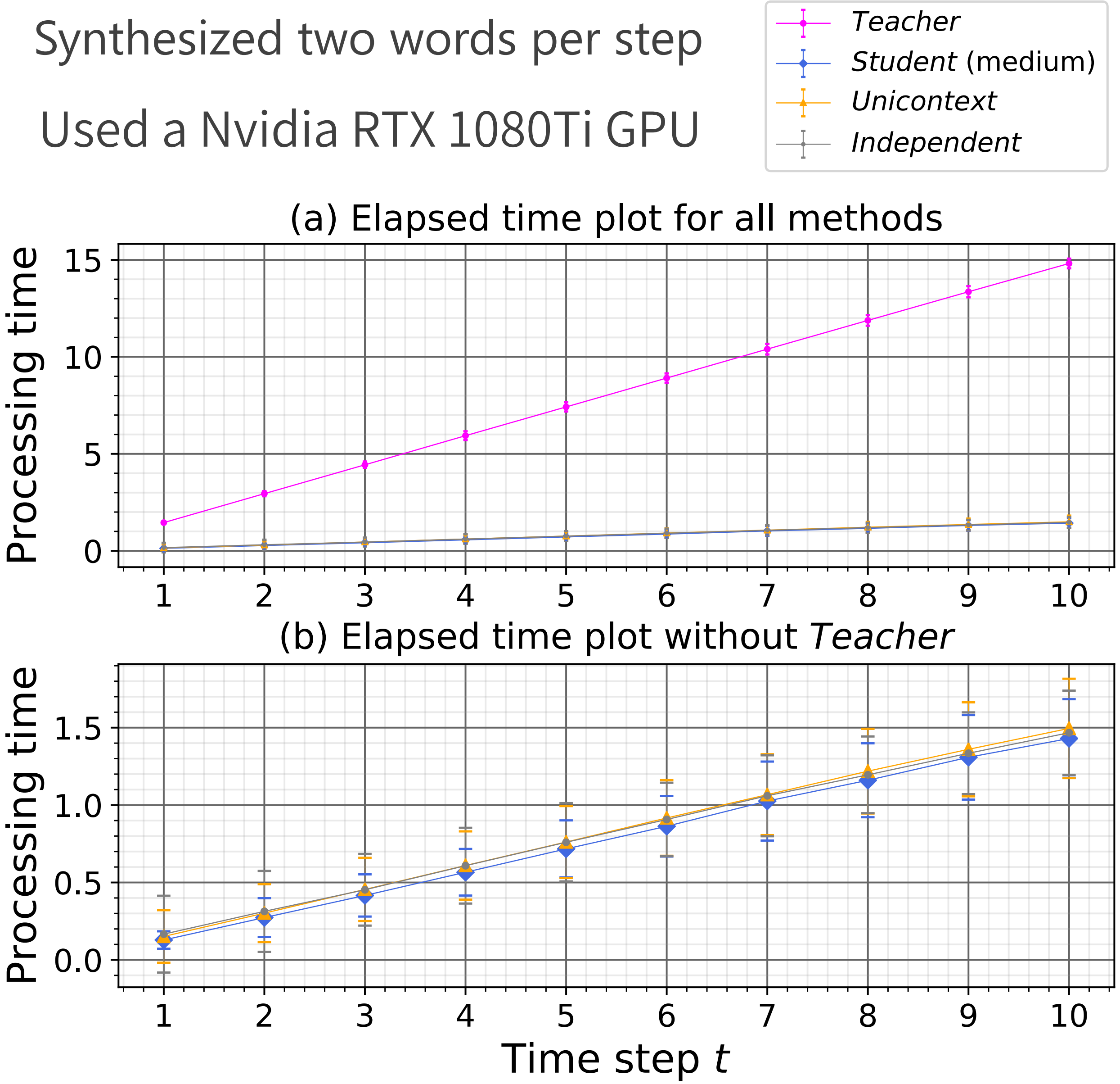
Average English speaker: **180** WPM
Teacher: **80** WPM
Student: **800** WPM

WPM: words per minute

Student achieves inference speed which can be available to **real-time application** while achieving **comparable quality** to Teacher

Synthesized two words per step

Used a Nvidia RTX 1080Ti GPU



Teacher [Saeki+, 2021] (80 WPM)

Ground-truth

Student w/o target loss (800 WPM)

Summary and Conclusion

Research goal

Low-latency and high-quality streaming TTS for real-time speech generation

Proposed method

Fast listen-while-predict framework that estimates future context with lightweight model

Knowledge distillation of context estimation model with GPT2 to single recurrent model

Evaluation results

Student model predicted effective contextual embedding for incremental TTS

Student model achieved **comparable synthetic speech quality to Teacher model**

Student model achieved much **faster speaking speed than human English speaker**

Future work

Further improving synthetic speech quality for equivalent quality to sentence-level TTS