Sec 3.2.3

Knowledge Distillation

Knowledge distillation

Knowledge distillation

E2E SLT (Student)

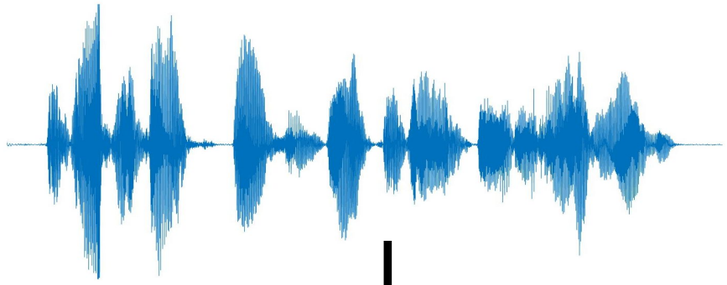
Knowledge distillation

Knowledge distillation

E2E SLT (Student)

MT (Teacher)

Knowledge distillation

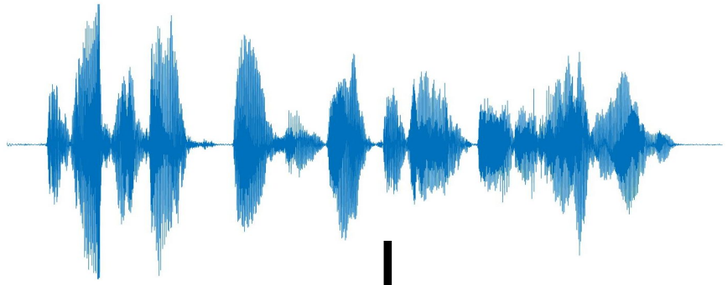


E2E SLT (Student)

This is the transcript of the speech

MT (Teacher)

Knowledge distillation



This is the transcript of the speech

E2E SLT (Student)

MT (Teacher)

How can the student learn from the teacher?

Knowledge Distillation

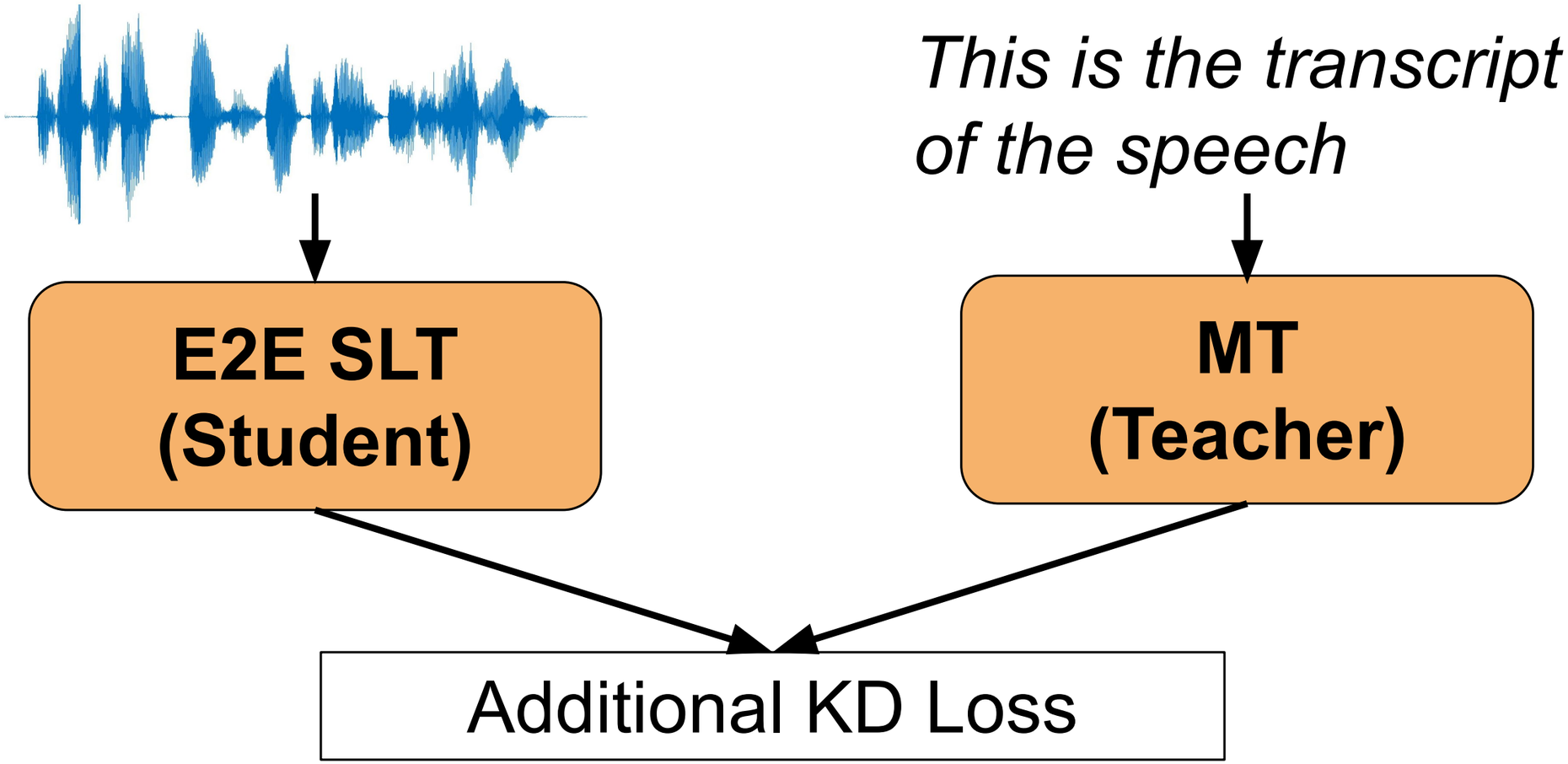
Knowledge distillation for sequences (Kim and Rush, 2016)

- Word-Level KD
- Sequence KD
- Sequence Interpolation KD

- Requirements:
  - ASR data
  - Pre-trained MT system

Word-Level KD

- Proposed by Liu et al. (2019)

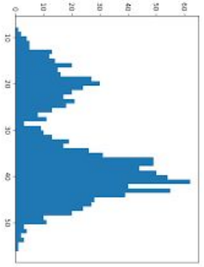


Word-Level KD

E2E SLT (Student)

MT (Teacher)

During training

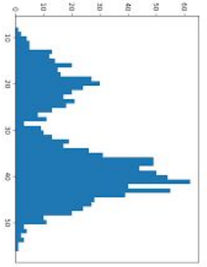


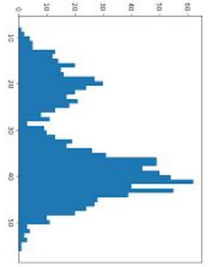
Word-Level KD

E2E SLT (Student)

MT (Teacher)

During training



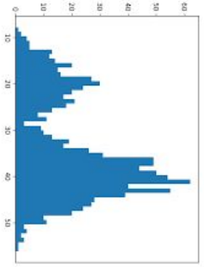


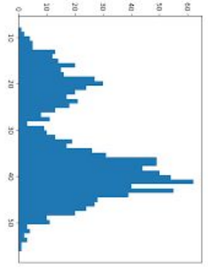
Word-Level KD

E2E SLT (Student)

MT (Teacher)

During training





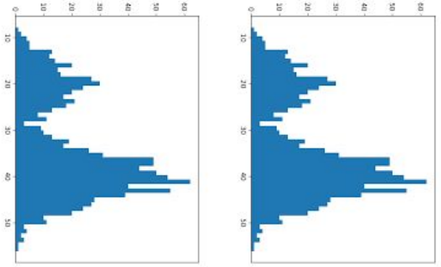
$KL(ST_1, MT_1)$

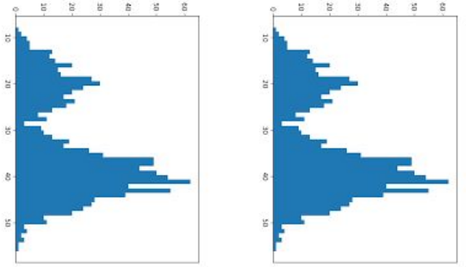
Word-Level KD

E2E SLT (Student)

MT (Teacher)

During training





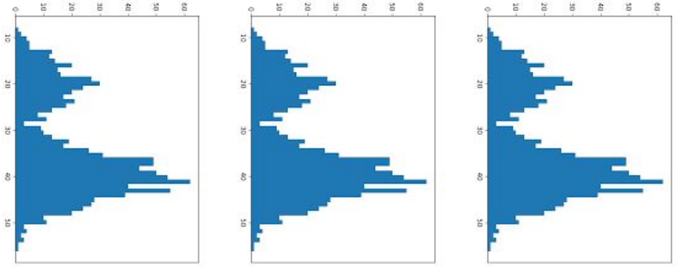
$KL(ST_1, MT_1)$ + $KL(ST_2, MT_2)$

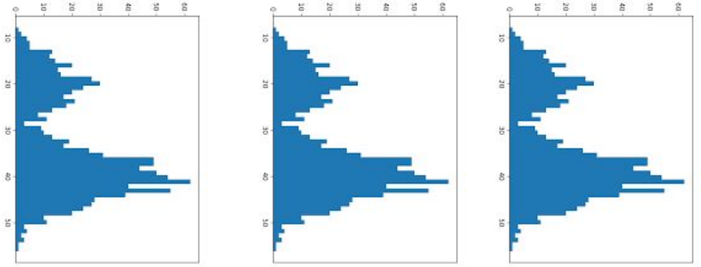
Word-Level KD

E2E SLT (Student)

MT (Teacher)

During training





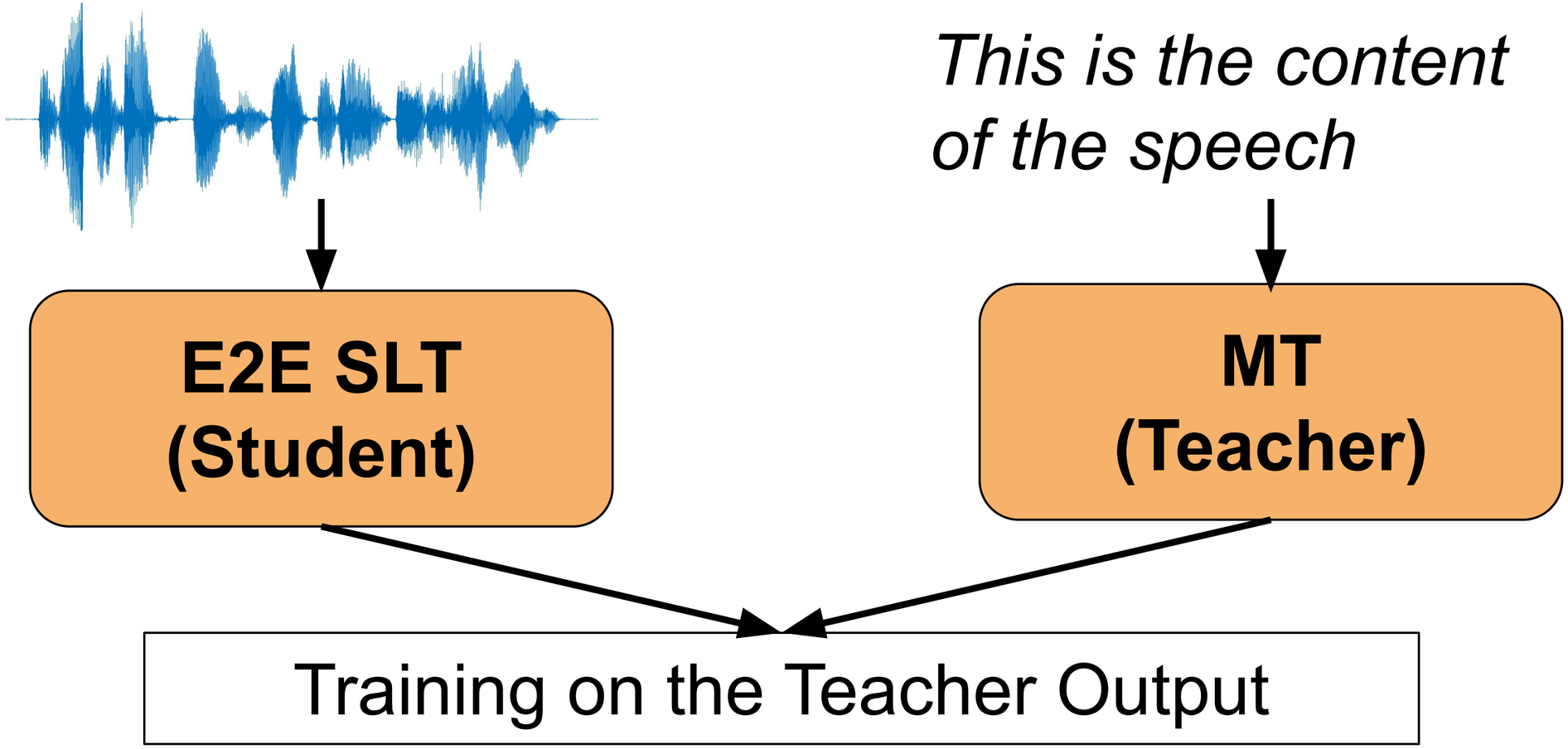
$KL(ST_1, MT_1)$ + $KL(ST_2, MT_2)$ + ...

Word-Level KD

- Training with SLT and KD losses
- Goal:
  - matching the output of SLT ground-truth
  - matching also the output probabilities of teacher model

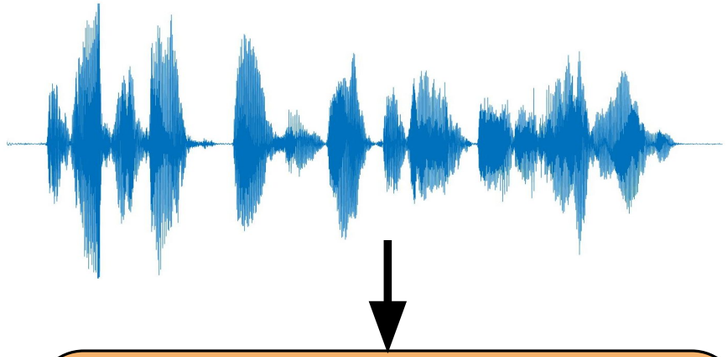
Sequence Level KD (Seq-KD)

- The output of the teacher is used as reference



Sequence Level KD (Seq-KD)

- The output of the teacher is used as reference



E2E SLT (Student)

This is the content of the speech

MT (Teacher)

Questo e' il contenuto del discorso

Sequence Level KD (Seq-KD)

- The output of the teacher is used as reference

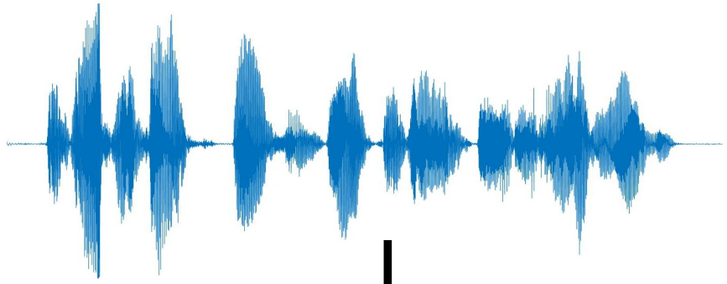
E2E SLT (Student)

MT (Teacher)

Questo e’ il contenuto del discorso

Sequence Interpolation (Seq-Inter)

- The n-bests of the teacher are rescored



This is the content of the speech

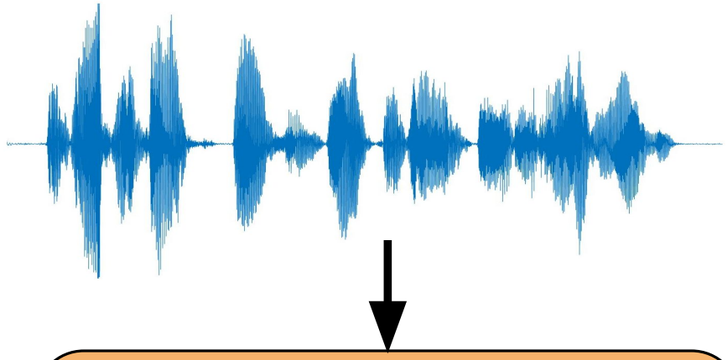
E2E SLT (Student)

MT (Teacher)

Training on the Rescored Teacher Output

Sequence Interpolation (Seq-Inter)

- The n-bests of the teacher are rescored



E2E SLT (Student)

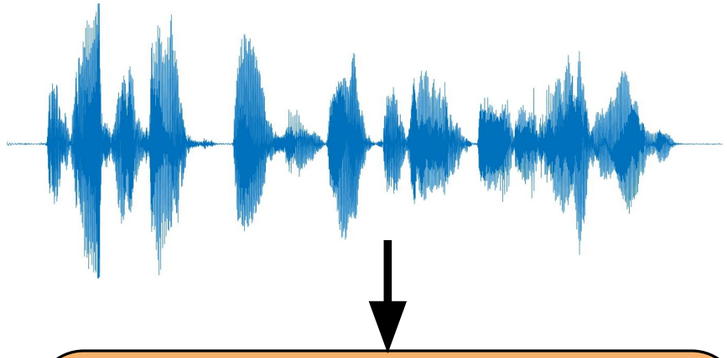
This is the content of the speech

MT (Teacher)

Questo e' il contenuto del discorso
Questo e' il contenuto dell'audio
Questo e' il contenuto

Sequence Interpolation (Seq-Inter)

- The n-bests of the teacher are rescored



E2E SLT (Student)

This is the content of the speech

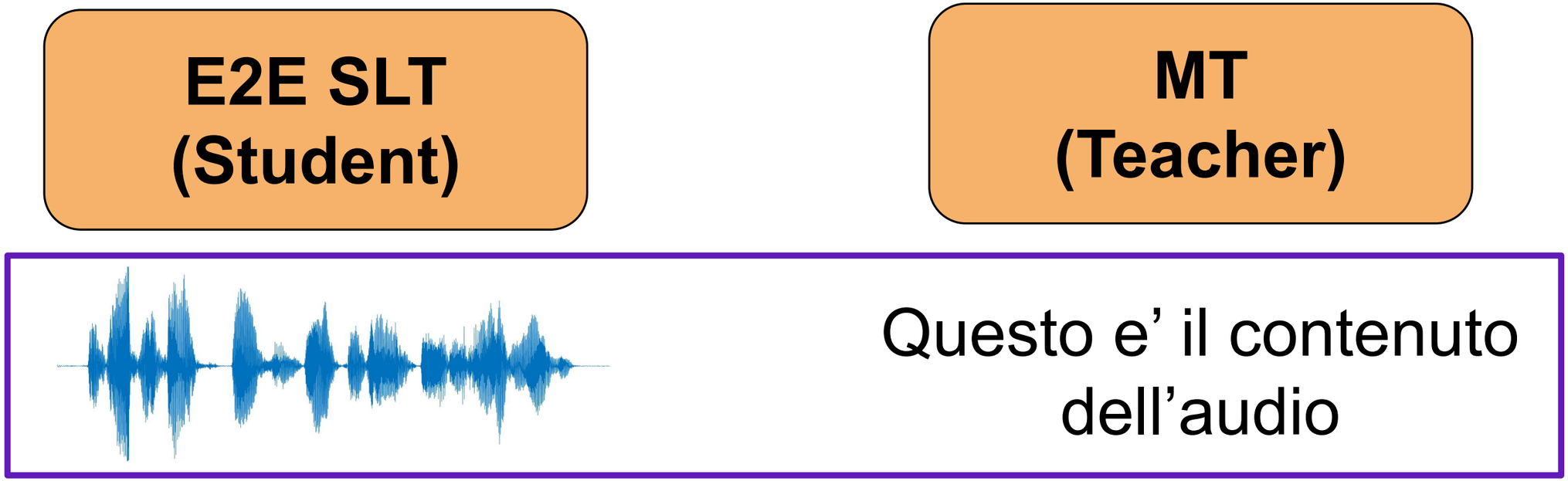
MT (Teacher)

Questo e' il contenuto dell'audio
Questo e' il contenuto del discorso
Questo e' il contenuto

Re-ranked n-best

Sequence Interpolation (Seq-Inter)

- The n-bests of the teacher are rescored



Sequence Interpolation (Seq-Inter)

How to rescore:

- BLEU using SLT data for which there is the reference
- Other methods: e.g. quality estimation (using ASR data)

Sequence Interpolation (Seq-Inter)

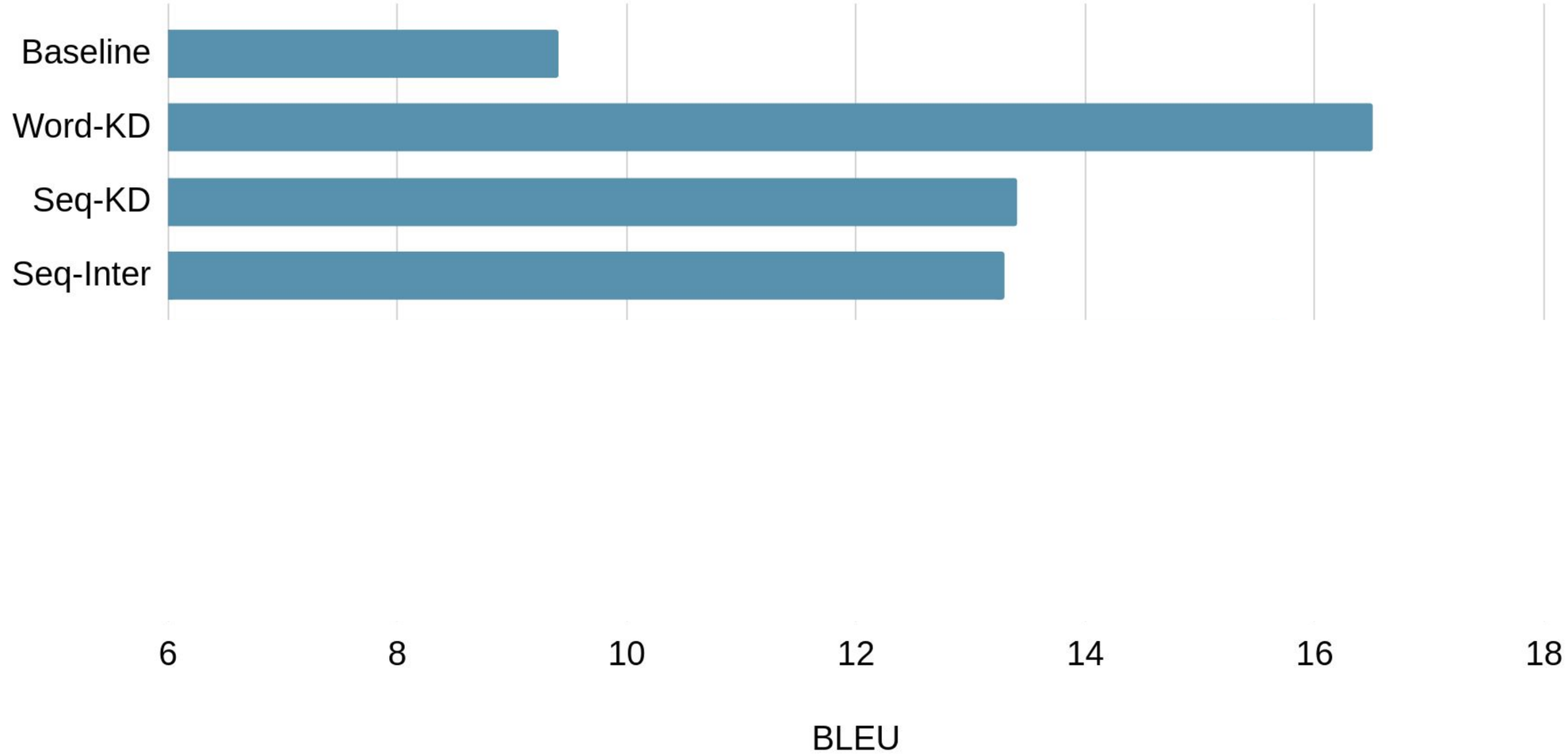
How to rescore:

- BLEU using SLT data for which there is the reference
- Other methods: e.g. quality estimation (using ASR data)

Goal:

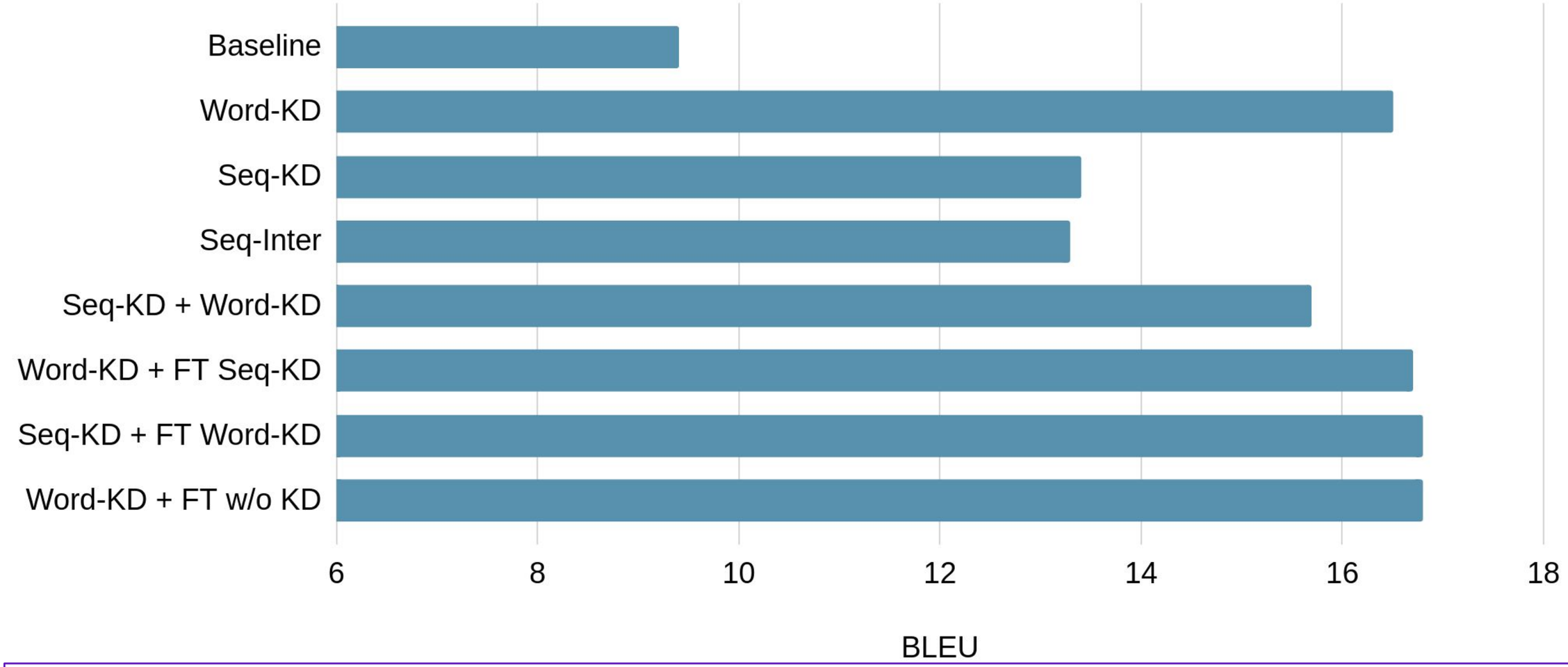
- To add knowledge from the teacher
- To reduce the lexical variability in the data (MT outputs have less variability)

KD Methods (Gaido et al., 2020)



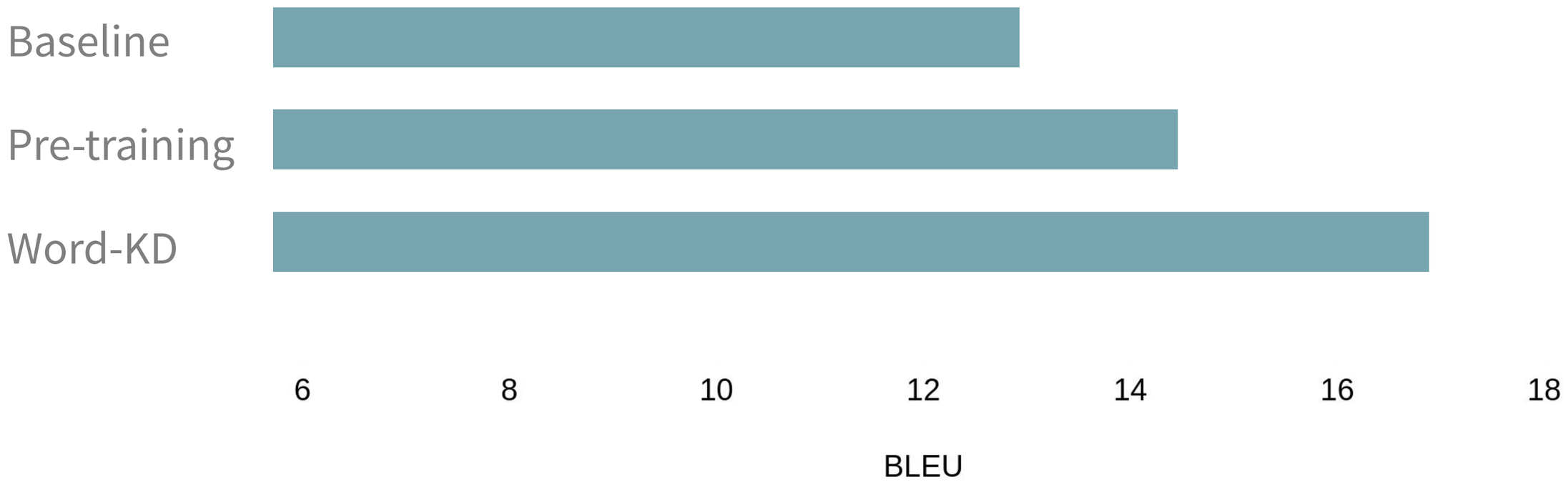
Word KD works the best

KD Methods (Gaido et al., 2020)



Word KD with a fine-tuning slightly improves over word KD

Pre-training vs KD (Liu et al., 2019)



KD outperforms pre-training