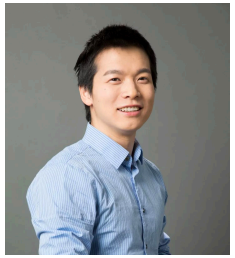
30 August – 3 September 2021
INTERSPEECH 2021 BRNO | CZECHIA
Speech everywhere!

# End-to-end Speech Translation via Cross-modal Progressive Training

***Rong Ye, Mingxuan Wang, Lei Li***





**ByteDance AI Lab**

Paper:

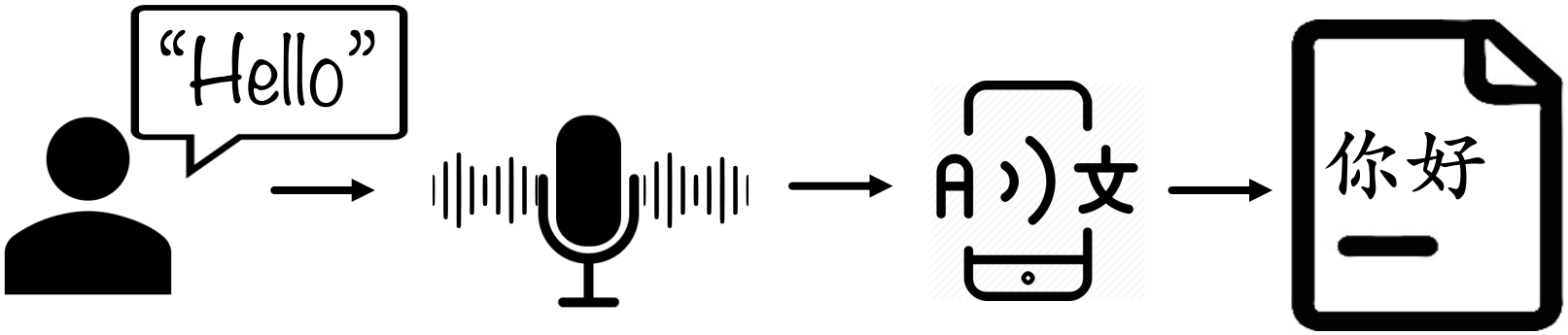
https://arxiv.org/abs/2104.10380

Code:

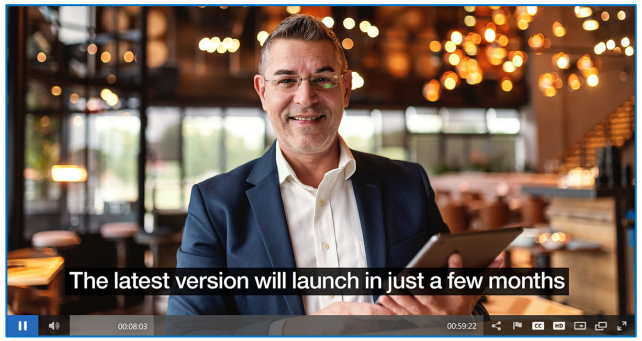
https://github.com/ReneeYe/XSTNet

# Speech-to-Text Translation (ST)

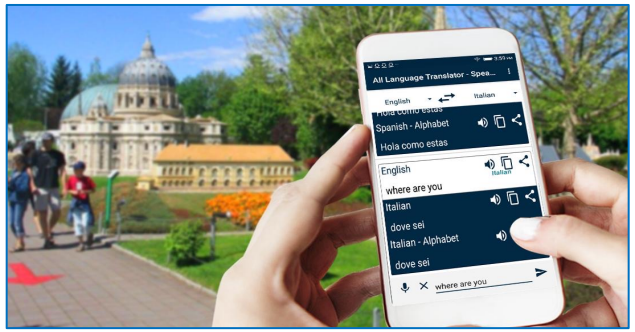
- Task: Source language ***speech(audio)*** → Target lang ***text***




- Wide applications of ST



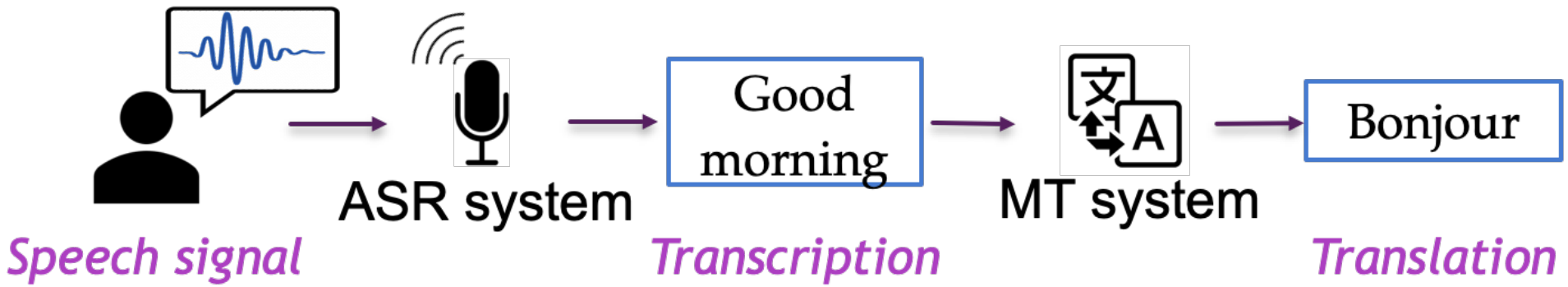

Foreign Media




Tourism

Global Conferences

# Cascaded ST System

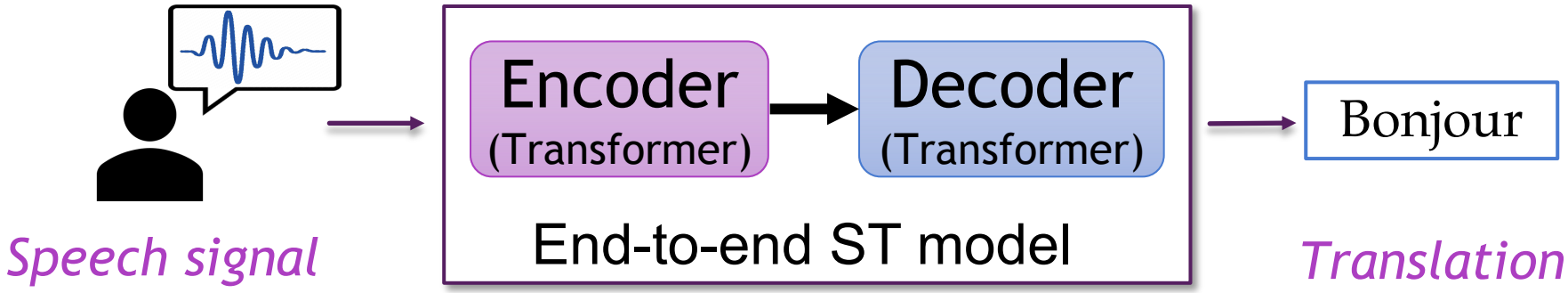



## - Challenges:

1. **Computationally inefficient**
2. **Error propagation**: *Wrong transcription → Wrong translation*

*do at this and see if it works for you* → 这样做，看看它是否对你有用
*duet this and see if it works for you* → 二重奏一下，看看它是否对你有用

# End-to-end ST Model




- Single model to produce text translation from the speech[1]
- **Classic models**: LSTM, Speech Transformer
- **Advantage:**
  - Reduced latency, simpler deployment
  - Reliving error propagation

[1] Bérard et al., Listen and translate: A proof of concept for end-to-end speech-to-text translation. 2016

# Challenges of End-to-end ST Model

- **Data scarcity** - lack of large parallel corpus

  `<speech, transcript_text, translate_text>`
- **Modality disparity** between speech and text
- Require low **latency** for product serving

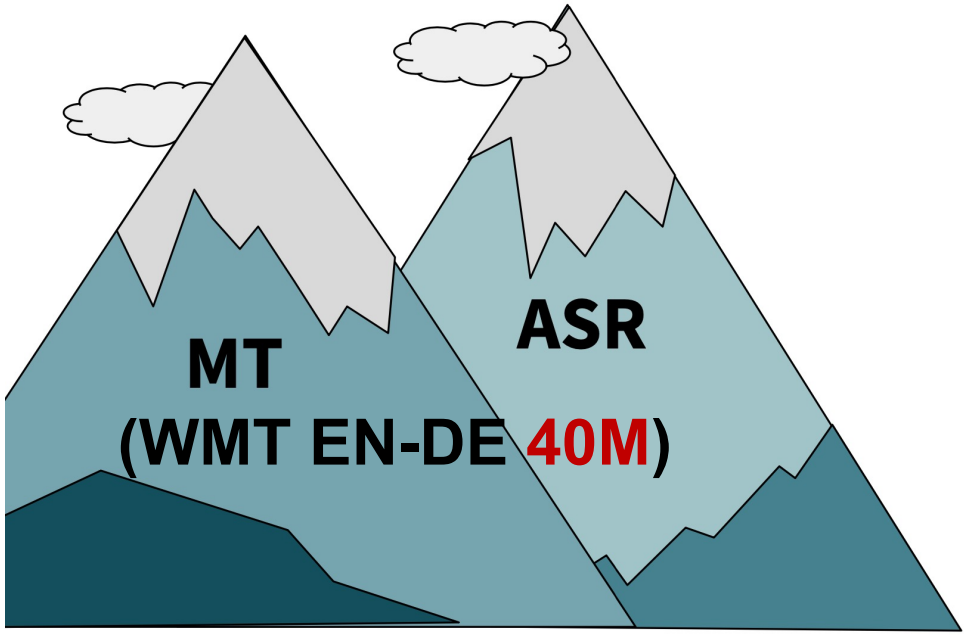
…

# Challenges of End-to-end ST Model

☹ **Data scarcity** - lack of large parallel corpus

`<speech, transcript_text, translate_text>`

- **Modality disparity** between speech and text
- Require low **latency** for product serving



# Two motivations to solve data scarcity

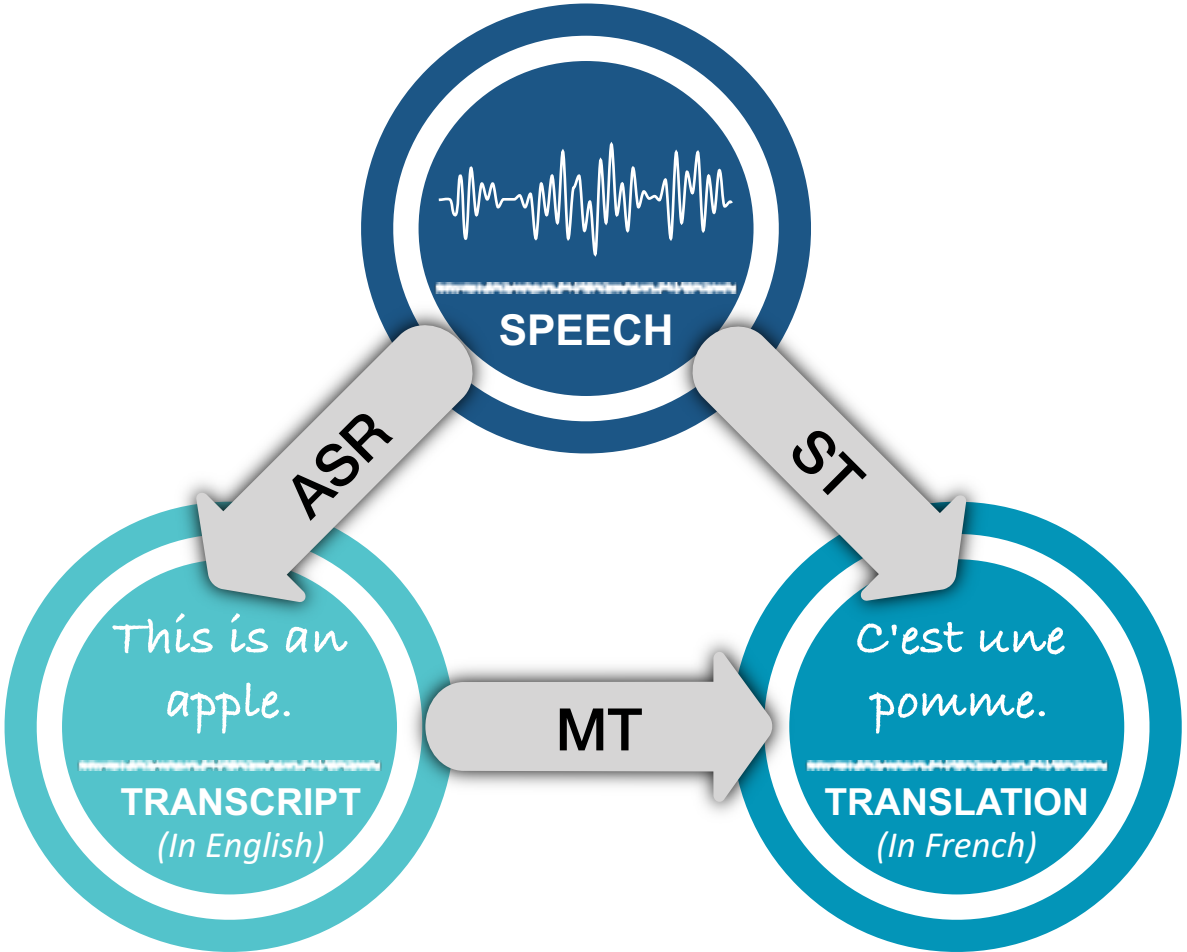
1. Fully utilize the existing
   <speech, transcript, translation> data.

2. Introduce large-scale MT data.

# Motivation1: Fully Utilize the ST triple data

ST triple data

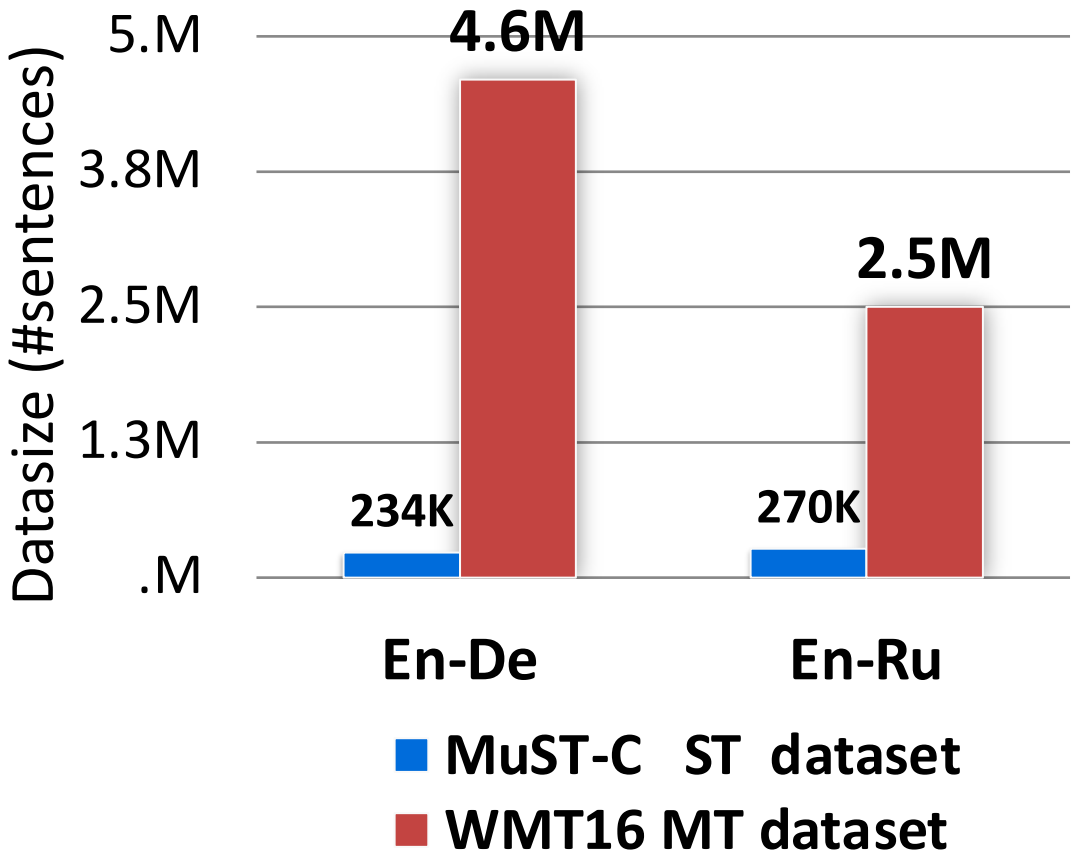
<Speech, Transcript, Translation> supervision.




Decomposed into three sub-tasks with parallel supervision, ST, ASR and MT.

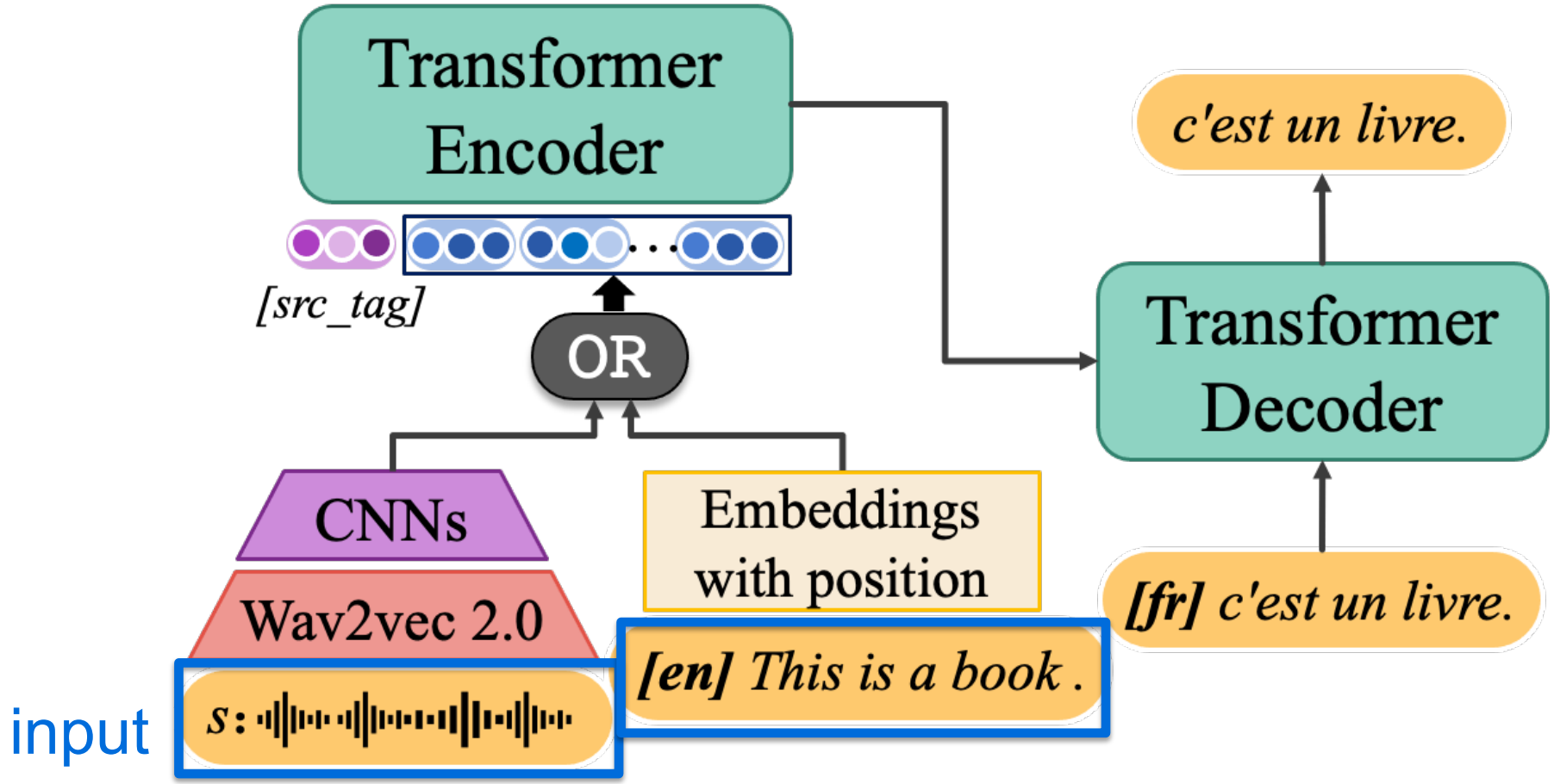
# Motivation 2: Using large-scale MT data

***Comparison of dataset size between ST and MT***



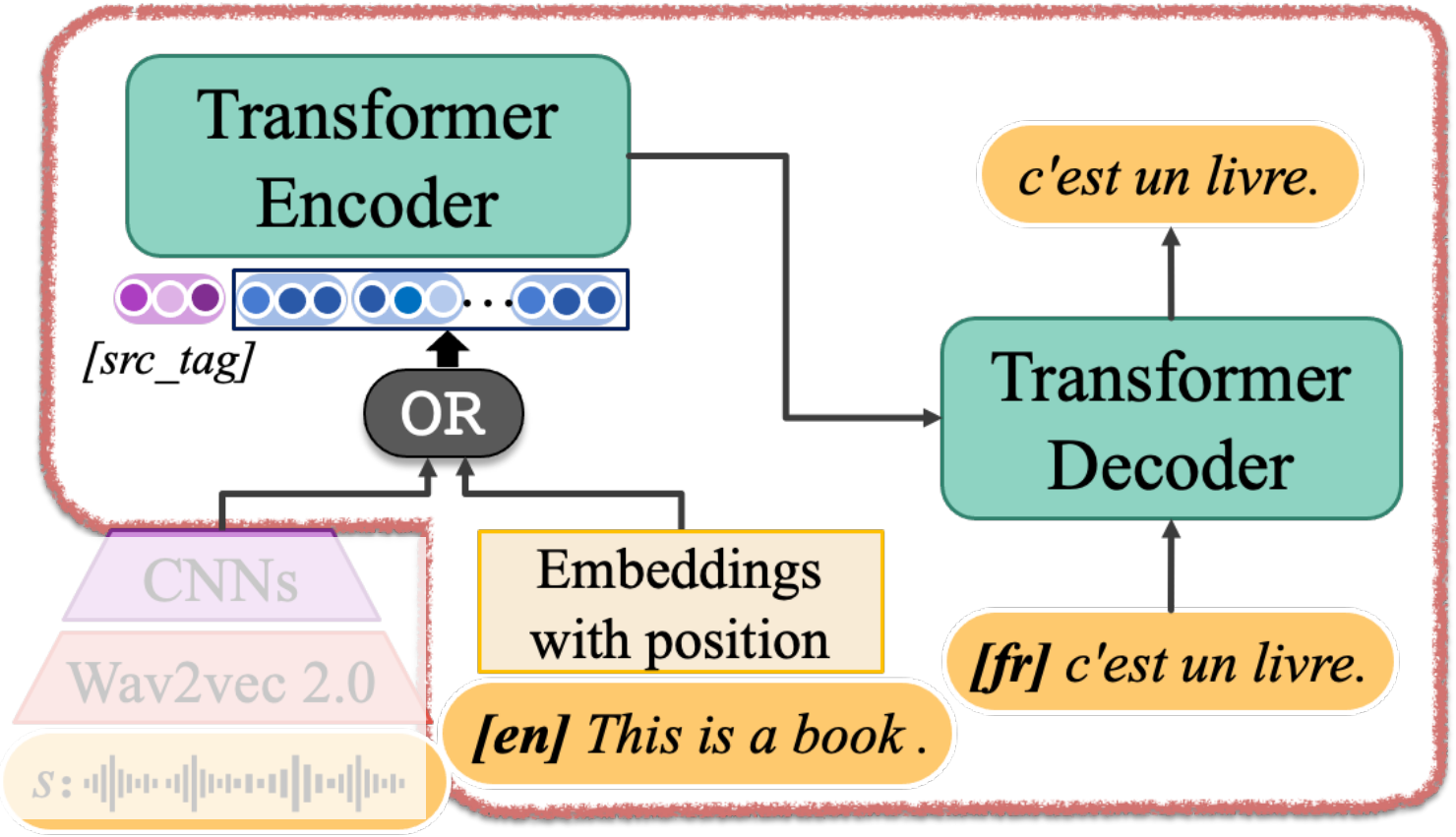

🤔 How to introduce MT data ***with much larger scale*** to improve ST performance?

# Cross Speech-Text Network (XSTNet)



# Supports to train MT data

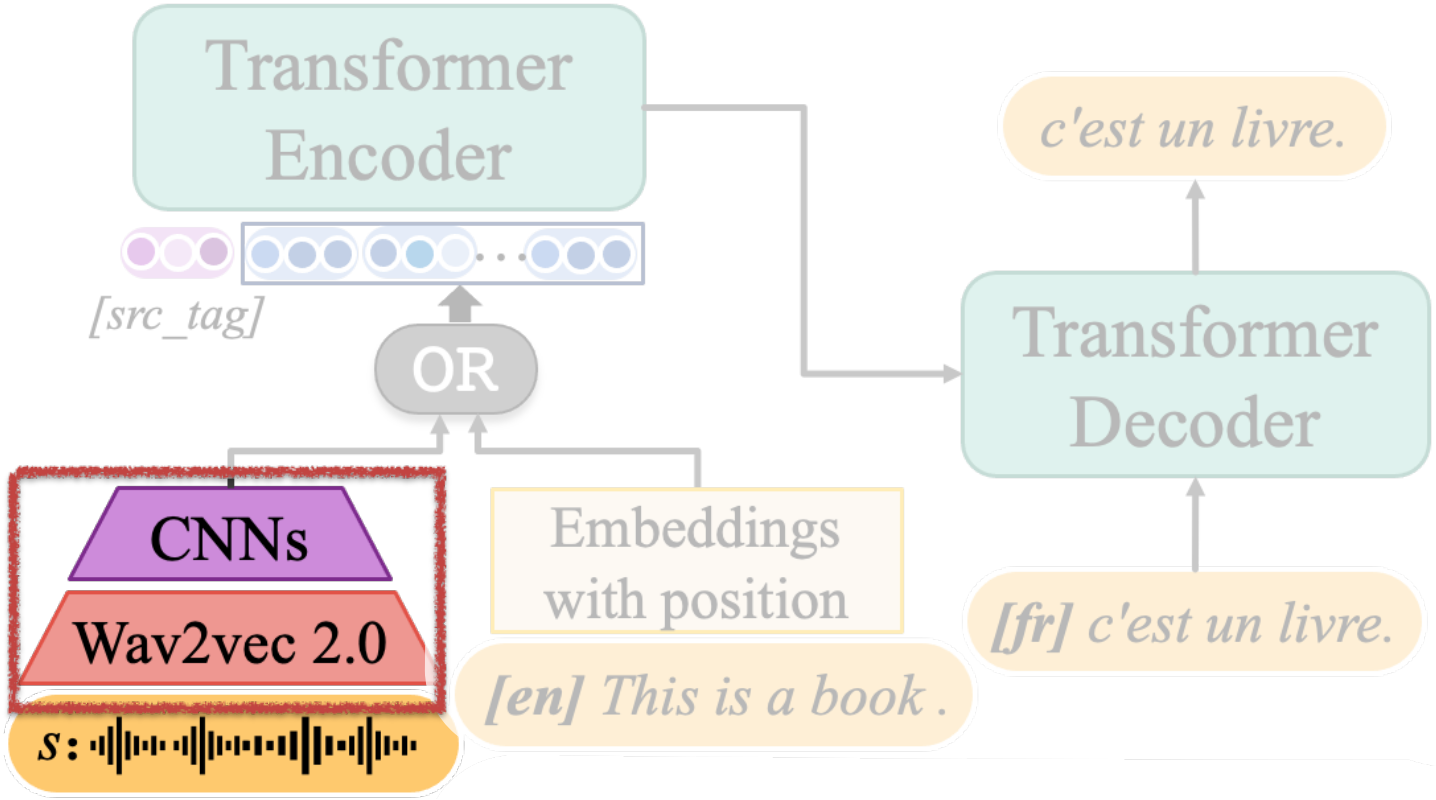
- Transformer MT model
- Add **more external MT data** to train Transformer encoder & decoder



# Supports inputs of two modalities

- Wav2vec2.0[1] as the acoustic encoder
- We add two convolution layers with 2-stride to shrink the length.

≈ Acoustic encoder

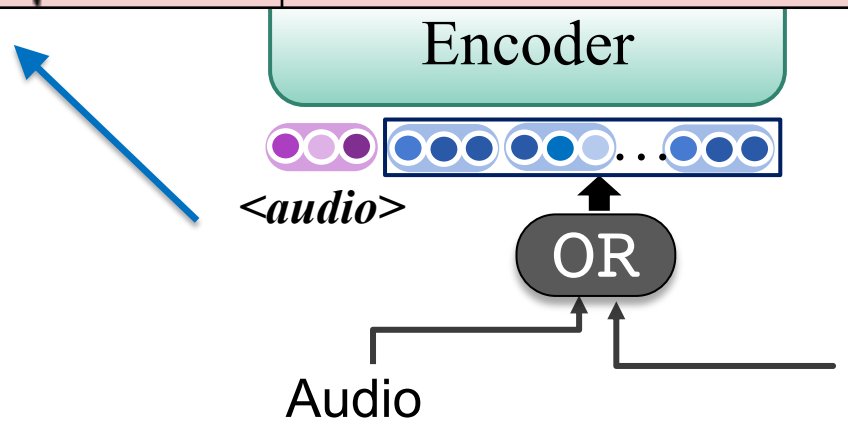



[1] wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020

# Language indicator strategy

- We use **language indicators** to distinguish different tasks.

| Tasks | Source input | Target output |
|---|---|---|
| MT | <en> This is a book. | <fr> c'est un livre. |
| ASR | <audio> | <en> This is a book. |
| ST | <audio> | <fr> c'est un livre. |



# Progressive Multi-task Training

## #1 Large-scale MT pre-training

Using **external MT** $D_{MT-ext}$

## #2 Multi-task Finetune

Using **(1) external MT** $D_{MT-ext}$

(2) $D_{ST}$ with <*speech, translation*>

(3) $D_{ASR}$ with <*speech, transcript*>

(4) $D_{MT}$ with <*transcript, translation*>

**Progressive:**

*Don't stop training* $D_{MT-ext}$

# XSTNet achieves State-of-the-art Performance

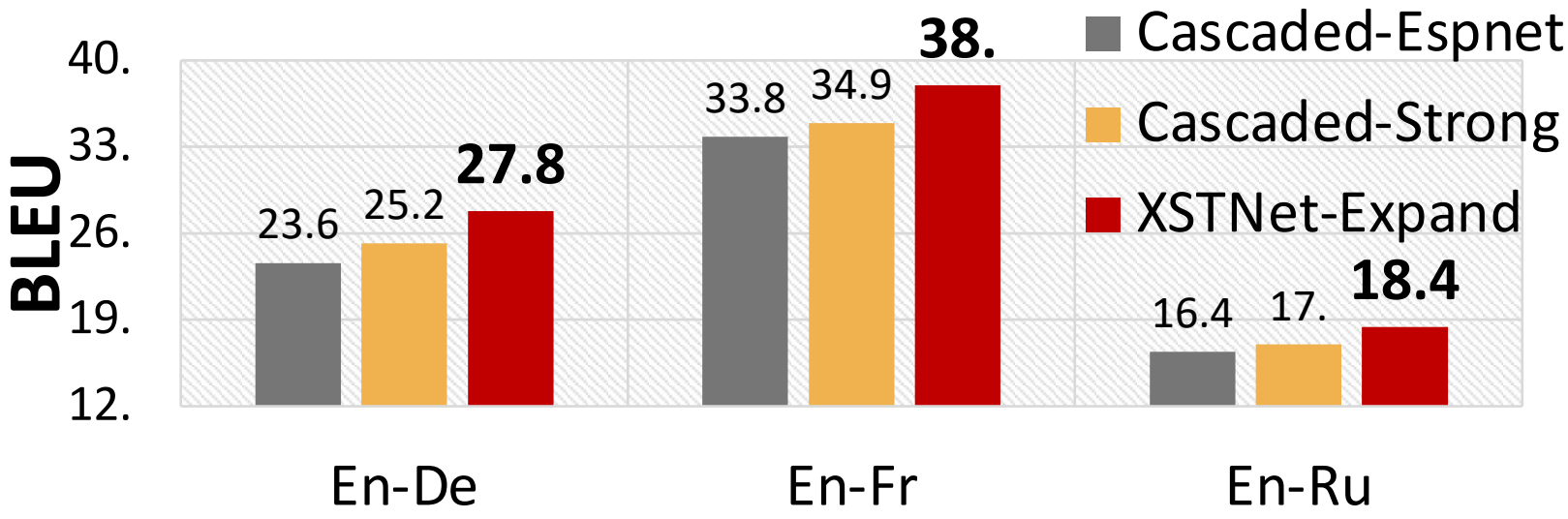
| Models | External Data | Pre-train Tasks | De | Es | Fr | It | Nl | Pt | Ro | Ru | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Transformer ST [13] | × | ASR | 22.8 | 27.4 | 33.3 | 22.9 | 27.2 | 28.7 | 22.2 | 15.1 | 24.9 |
| AFS [31] | × | × | 22.4 | 26.9 | 31.6 | 23.0 | 24.9 | 26.3 | 21.0 | 14.7 | 23.9 |
| Dual-Decoder Transf. [15] | × | × | 23.6 | 28.1 | 33.5 | 24.2 | 27.6 | 30.0 | 22.9 | 15.2 | 25.6 |
| Tang et al. [2] | MT | ASR, MT | 23.9 | 28.6 | 33.1 | - | - | - | - | - | - |
| FAT-ST (Big) [6] | ASR, MT, mono-data† | FAT-MLM | 25.5 | 30.8 | - | - | 30.1 | - | - | - | - |
| W-Transf. | audio-only* | SSL* | 23.6 | 28.4 | 34.6 | 24.0 | 29.0 | 29.6 | 22.4 | 14.4 | 25.7 |
| **XSTNet (Base)** | audio-only* | SSL* | 25.5 | 29.6 | 36.0 | 25.5 | 30.0 | 31.3 | 25.1 | 16.9 | 27.5 |
| **XSTNet (Expand)** | MT, audio-only* | SSL*, MT | **27.8**§ | **30.8** | **38.0** | **26.4** | **31.2** | **32.4** | **25.7** | **18.5** | **28.8** |

Table 1: *Performance (case-sensitive detokenized BLEU) on MuST-C test sets.* †*: "Mono-data" means audio-only data from Librispeech, Libri-Light, and text-only data from Europarl/Wiki Text; *: "Audio-only" data from LibriSpeech is used in the pre-training of wav2vec2.0-base module, and "SSL" means the self-supervised learning from unlabeled audio data.* §*uses OpenSubtitles as external MT data.*

**XSTNet (Base)**: Achieves the SOTA in the restricted setup

**XSTNet (Expand)**: Goes better by using extra MT data

# XSTNet better than cascaded ST! with a gain of 2.6 BLEU

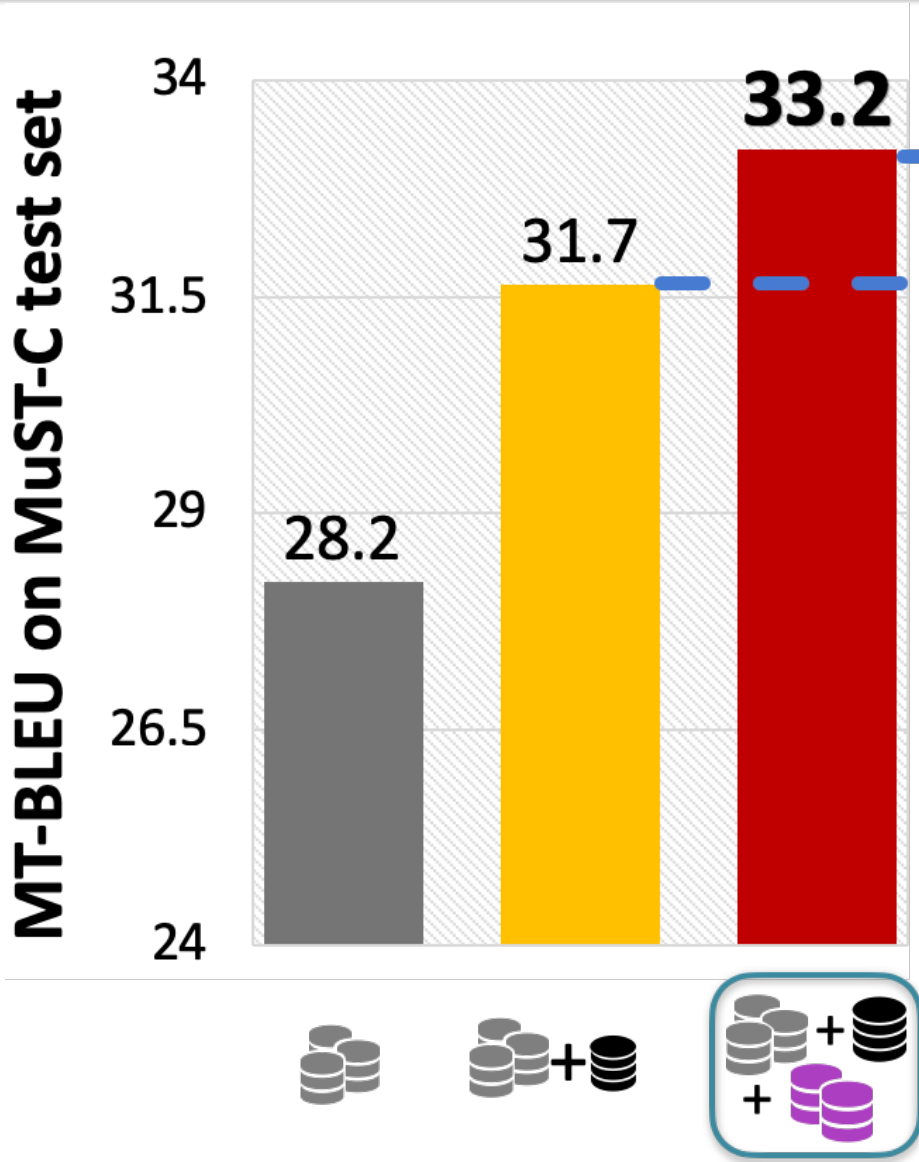



What is "Cascaded-Strong" system?

Strong ASR model + Large-scale MT data

| **Cascaded - Strong** | **Model** | **Training data** | Performance (En-De) |
|---|---|---|---|
| ASR | W2V2+ Transformer | MuST-C $D_{ASR}$ | WER=13.0 |
| MT | Transformer-base | WMT + MuST-C $D_{MT}$ | BLEU=31.7 |

# MT ⬆ from audio information




**+ 1.5 BLEU**

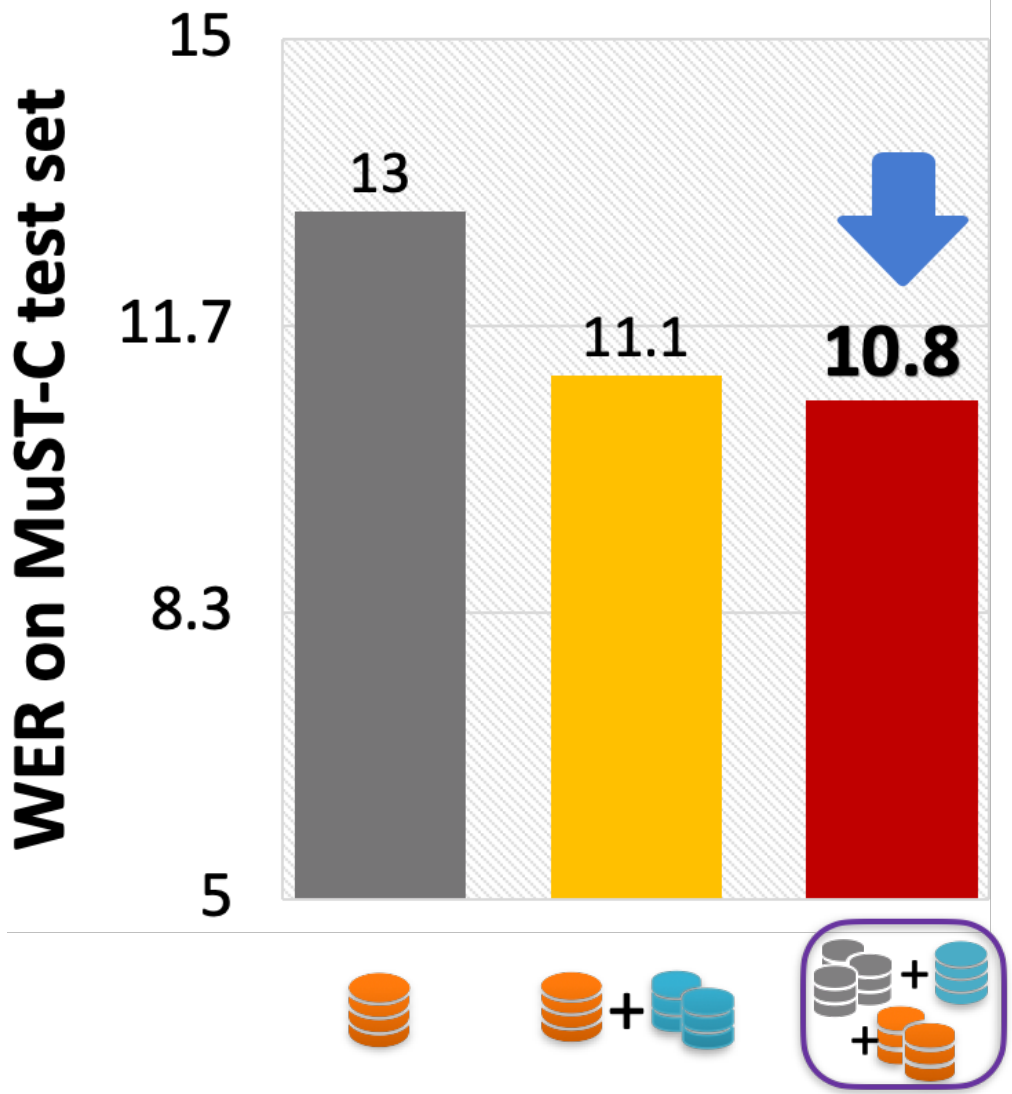
MT can be improved from audio information.

WMT

MuST-C $D_{MT}$

MuST-C $D_{ST} \cup D_{ASR}$

# ASR ⬆️ from text information




## - 2.2 WER

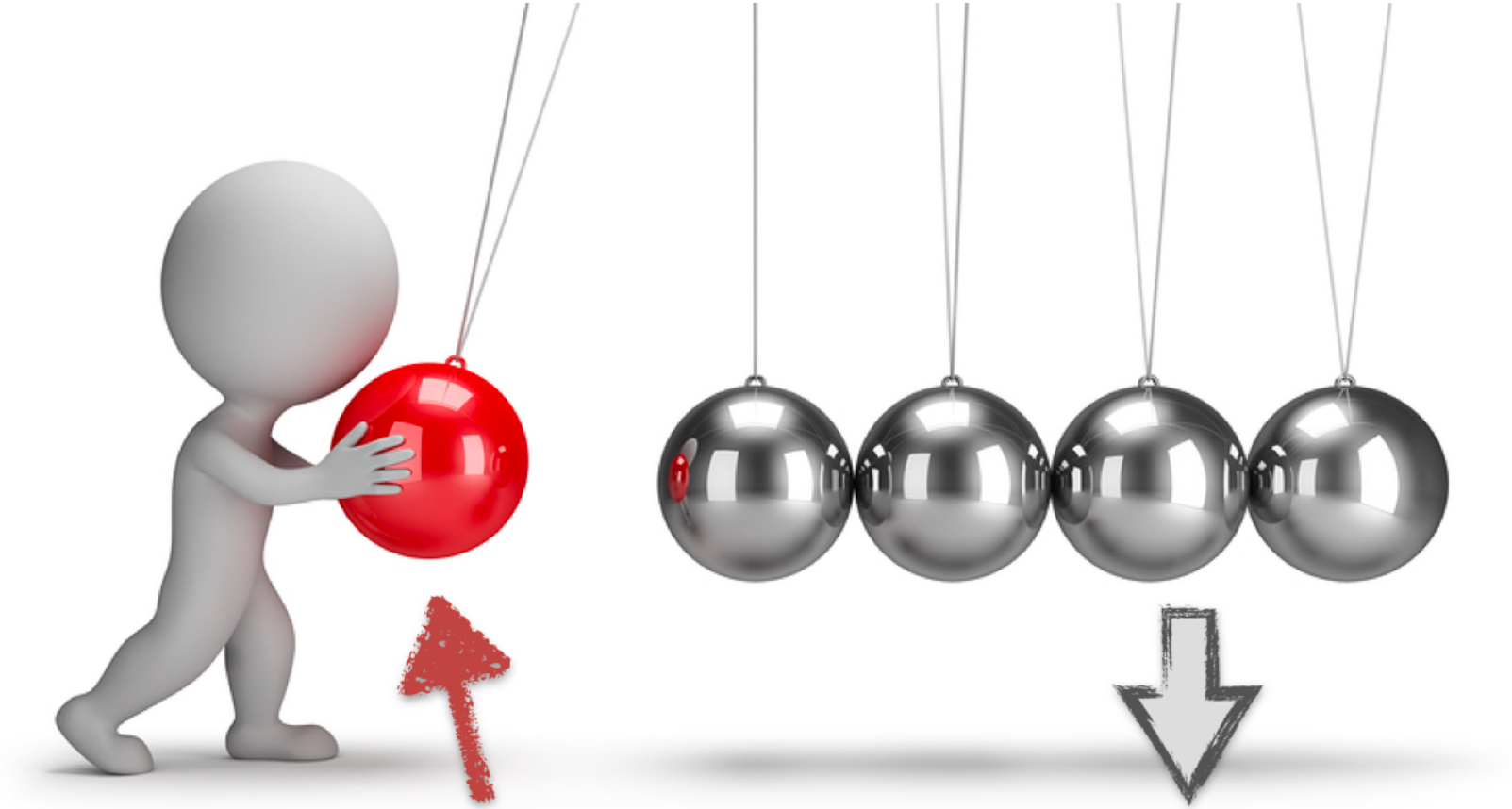
Multi-task training also benefits ASR.

WMT

MuST-C $D_{ASR}$

MuST-C $D_{ST} \cup D_{MT}$

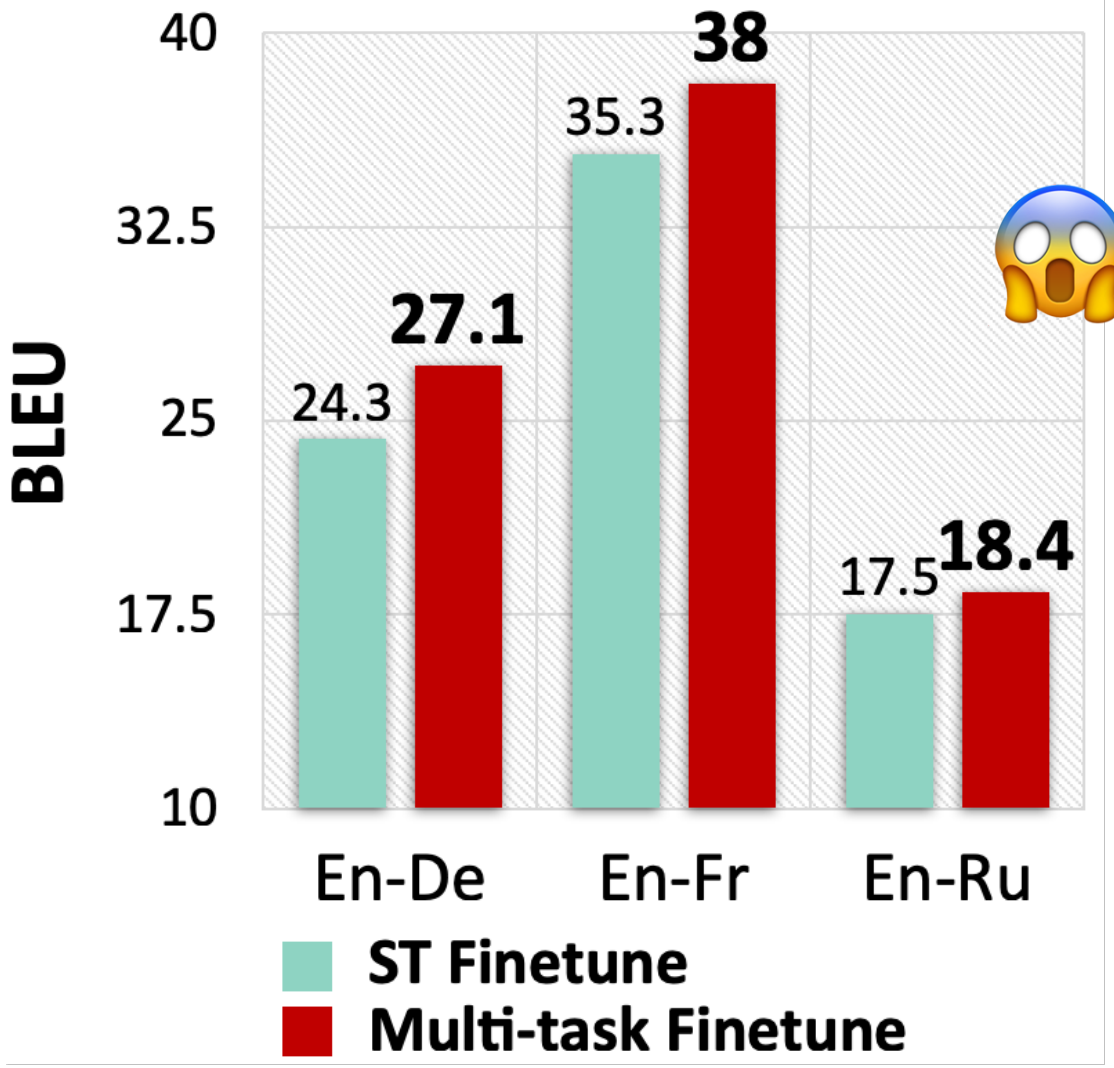
# Training strategy <u>really</u> matters!



Small training tricks

Large impact on ST performance

# Multi-task fine-tuning is MUCH better

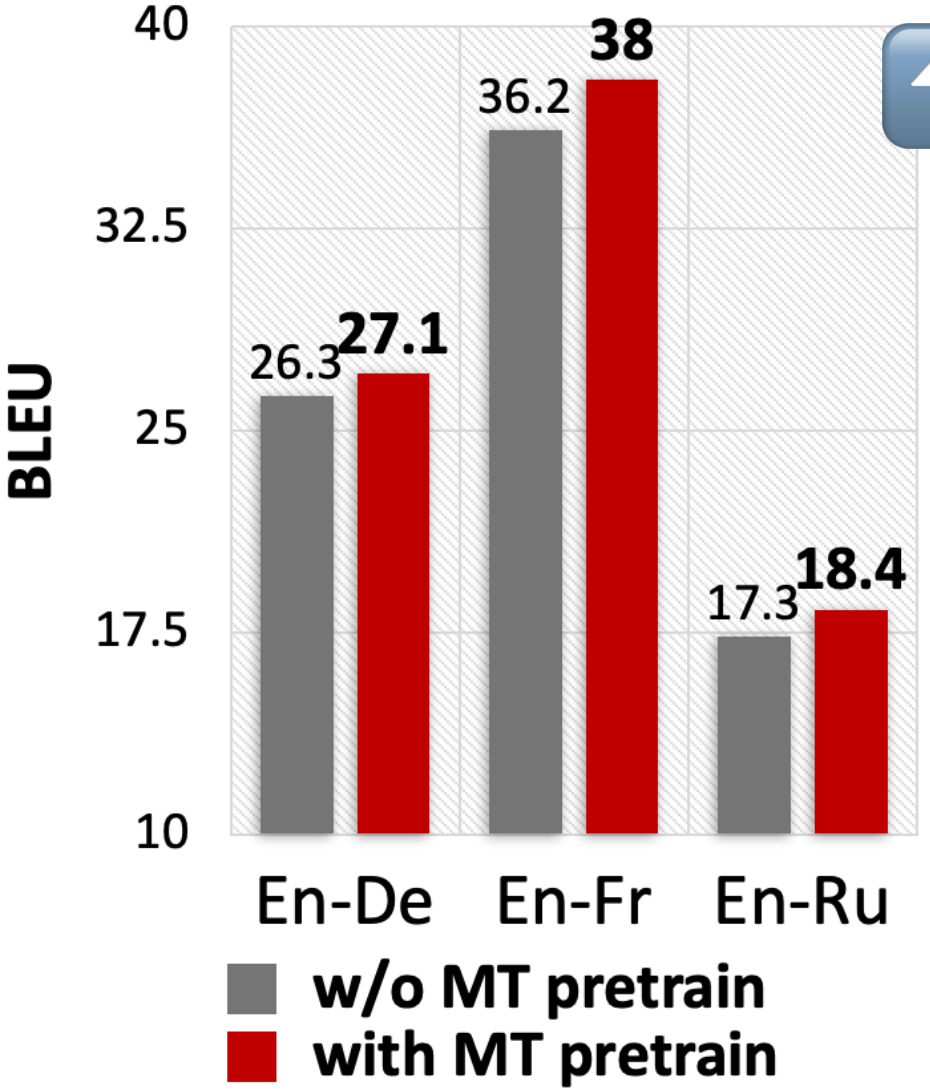





**2.1 BLEU** ⬆️

if we apply multi-task strategy in finetuning

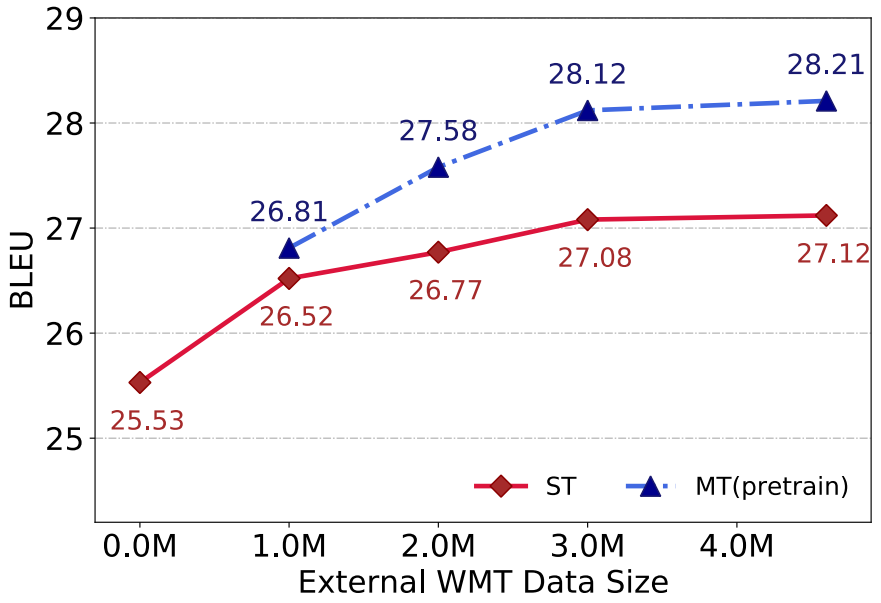
# MT pre-training is necessary



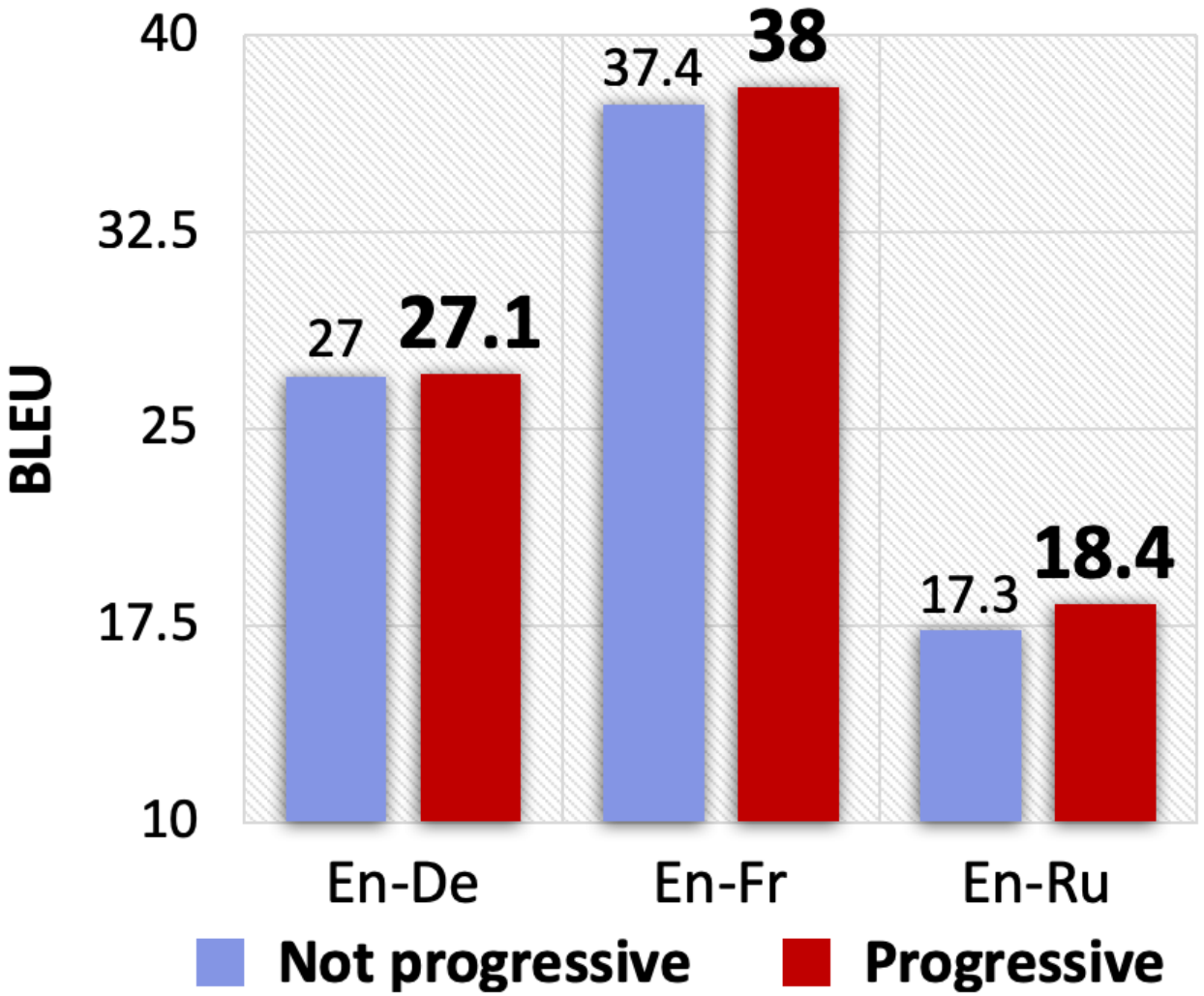

## 1.2 BLEU

due to the MT pre-train

And, the more extra MT data, the better.



# Don’t stop training the data in the previous MT pretrain stage

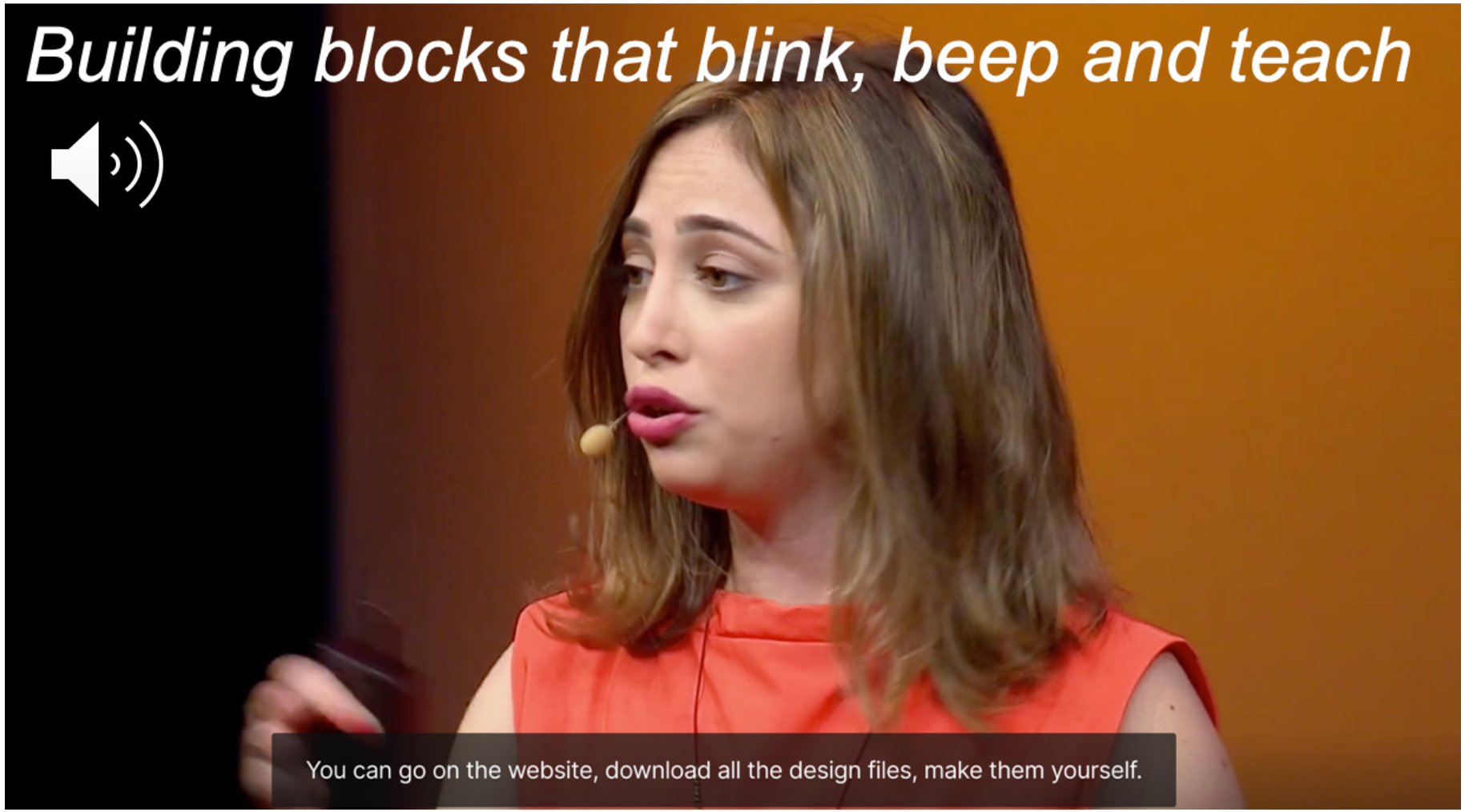



Continue using extra WMT data, and

 **0.3 BLEU**

# XSTNet can reduce the error propagation

Ayah Bdeir | TED 2012



# Errors in ASR transcript

- **Ground truth transcript:**

  *You can go on the website, download all the design files, make them yourself.*

- **Generated transcript:**

  *You can go on the website, download **other** design files, make them yourself.*

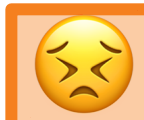
- **Translation generated**

  **XSTNet: (correct)**

  *Sie können auf der Website die Designdateien herunterladen, sie selbst herstellen.*

  **Cascade:**

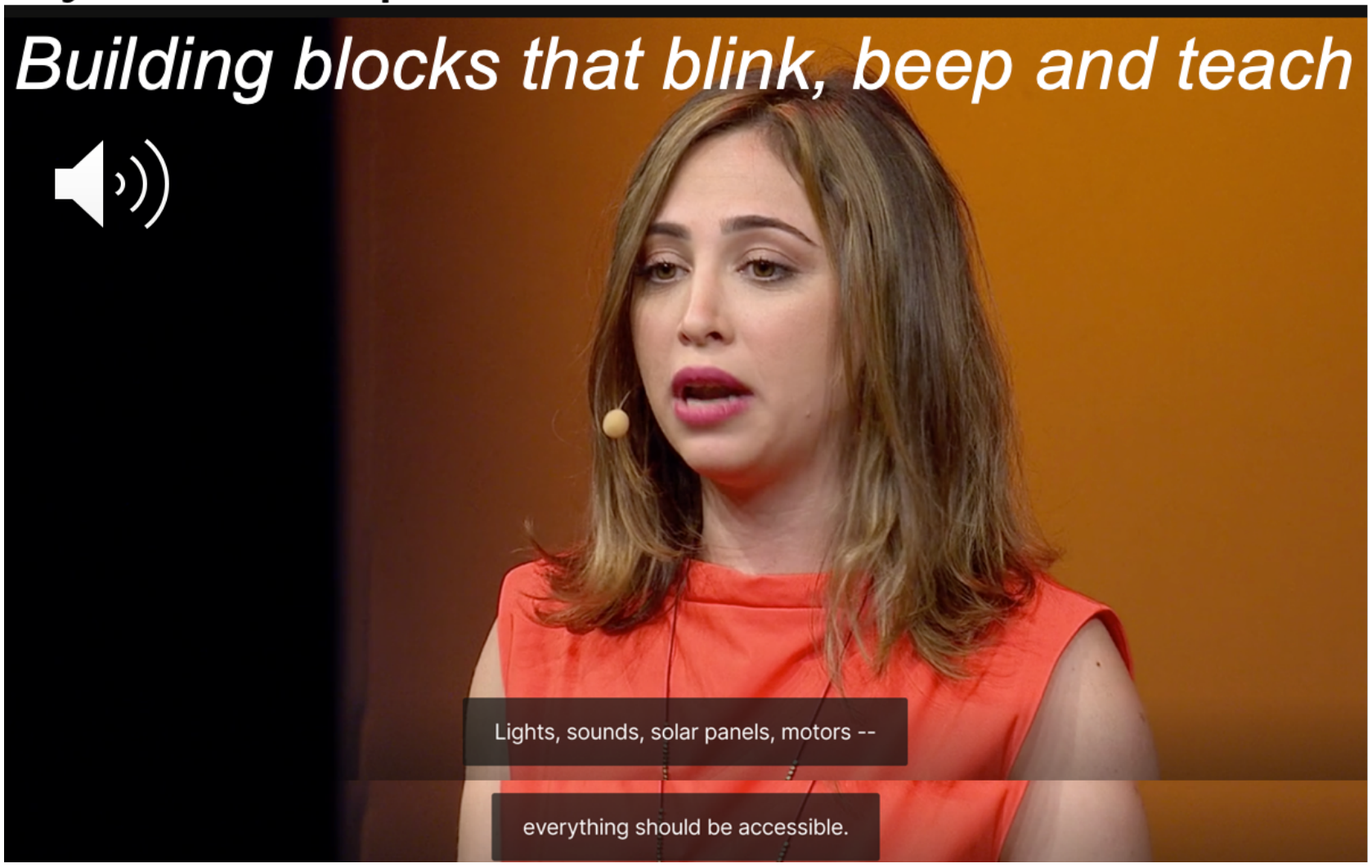
  *Sie können auf die Website gehen, **andere Designdateien herunterladen**, sich selbst machen.*



**andere** means “other”

# XSTNet reduces the error caused by ambiguous punctation

Ayah Bdeir | TED 2012



# Misinterpretation caused by ambiguous punctuation

- **Transcript generated:**

  Lights sounds solar panels motors everything should be accessible.

- **Translation generated**

  ✅ **XSTNet: (correct)**

  Licht, Geräusche, Solarkollektoren, Motoren — alles sollte zugänglich sein.

  ❌ **Cascade:**

  Licht **klingt** Solarpaneele, Motoren; alles sollte zugänglich sein.

😖 klingt is a verb, means “sound like”

# Takeaways of XSTNet

- XST-Net, an extremely concise model & has **excellent performance**, featuring:
  - Accept bi-modal inputs & external **MT** data
  - Multi-task: Jointly train ST, ASR and MT
- **Training process matters:**
  - **MT pre-training** is necessary
  - **Multi-task** > ST-only fine-tuning
  - Continue using external MT data, ie. **progressive** training

# THANKS

- Paper: https://arxiv.org/abs/2104.10380
- Code & models: https://github.com/ReneeYe/XSTNet
- Project Page: https://reneeye.github.io/projects/XSTNet *(keep updating)*
- E-mail: yerong@bytedance.com

PAPER CODE

## OTHER RELATED:

- NeurST Codebase: https://github.com/bytedance/neurst [ACL2021 demo]
-  VolcTrans: https://translate.volcengine.cn/