语音翻译：从前沿研究到产品创新

Speech Translation

李磊
字节跳动人工智能实验室
2021/6/6
Cross Language Barrier with Machine Translation

Foreign Media

Tourism

Global Conferences

International Trade

The latest version will launch in just a few months
Outline

1. Overview: ST Problem and Challenge
2. What is a better model for ST?
3. Better training strategy for ST?
4. New ST-powered Products
Speech-to-Text Translation (ST)

- source language *speech(audio) ➔ target lang* text

**Application Type**
- (Non-streaming) ST
- Streaming ST

**System**
- Cascaded ST
- End-to-end ST
Cascaded ST System

- Challenges:
  1. Computationally inefficient
  2. Error propagation: Wrong transcription ? Wrong translation

*do at this* and see if it works for you
*duet this* and see if it works for you
• Single model to produce text translation from speech
• Popular model: Encoder-Decoder architecture (e.g. Transformer)
• Advantage:
  – Reduced latency, simpler deployment
  – Avoid error propagation

Basic Speech Translation Architecture (Same as MT)

Transformer-based: N-layer encoder, M-layer decoder
Challenge

- Data scarcity - lack of large parallel corpus
- Modality disparity between audio and text
- Require low latency for product serving
Approaches for End-to-end ST

• Model
  – Better Decoder: COSTT [AAAI 2021b]

• Training technique
  – Audio pre-training: Wave2Vec2.0 [Baevski et al 2021]
  – Progressive multi-task training: XSTNet [Interspeech 2021]

• Speed-up Inference (not in this talk)
  – Parallel Decoding: GLAT [ACL 2021b]
  – GPU optimization: LightSeq [NAACL2021]
Listen, Understand and Translate: Triple Supervision Decouples End-to-end Speech-to-text Translation

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, Lei Li
1. A single encoder is hard to capture the representation of audio for the translation.  
2. Limited in utilizing the information of “transcription” in the training.
Motivation: Mimic human’s behavior

Question: How human translate?

“Listen-Understand-Translate” (LUT) model based motivated by human’s behavior
Motivation of Better Encoding

**Drawback 1:** A single encoder is not enough.

**Idea 1:** Introduce a semantic encoder

**Drawback 2:** Limit in using “transcript” info.

**Idea 2:** Utilizing the pre-trained representation (e.g. BERT) of the “transcript” to learn the semantic feature.
LUT for End-to-end ST

Training data: triples of

\(<\text{speech}, \text{transcript}\_\text{text}, \text{translate}\_\text{text}>\)

Transcript \((z)\):
“Good morning”

Input \((x)\):
Log-mel \(f_{\text{bank}}\) feature

CTC loss

Translation \((y)\):
“Bonjour”

\(\text{CE loss}\)

\(\text{Distance loss}\)

\(\text{BERT representation}\)

Acoustic Encoder (Listen)

Translation Decoder (Translate)

Semantic Encoder (Understand)

Learning Shared Semantic Space for Speech-to-Text Translation

Chi Han, Mingxuan Wang, Heng Ji, Lei Li

Code: https://github.com/Glaciohound/Chimera-ST
Speech and text interfere with each other in brain[1]  

Convergence sites of speech (blue) and text (yellow)  

Idea: Bridging the Speech-Text modality gap with Shared Semantic Representation

ST triple data:

<speech, transcript_text, translate_text>

Text Input
“everyone loves cats”

Audio Input

Translation
“tut le monde aime les chats”
Chimera Model for ST

Training with auxiliary objectives: ST + MT + Contrastive loss
Benefit: able to exploit large external MT data

Learning Shared Semantic Space for Speech-to-Text Translation Listen [Chi Han, Mingxuan Wang, Heng Ji, Lei Li, Findings of ACL 2021]
Chimera achieves the best (so far) BLEU on all languages in MuST-C

<table>
<thead>
<tr>
<th>Model</th>
<th>External Data</th>
<th>Speech</th>
<th>ASR</th>
<th>MT</th>
<th>MuST-C</th>
<th>EN-X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>EN-DE</td>
<td>EN-FR</td>
</tr>
<tr>
<td>FairSeq ST †</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>22.7</td>
<td>32.9</td>
</tr>
<tr>
<td>Espnet ST ‡</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>22.9</td>
<td>32.8</td>
</tr>
<tr>
<td>AFS *</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>22.4</td>
<td>31.6</td>
</tr>
<tr>
<td>Dual-Decoder ◊</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>23.6</td>
<td>33.5</td>
</tr>
<tr>
<td>STATST ‡</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>23.1</td>
<td>-</td>
</tr>
<tr>
<td>MAML b</td>
<td></td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>22.1</td>
<td>34.1</td>
</tr>
<tr>
<td>Self-Training ◎</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td></td>
<td>25.2</td>
<td>34.5</td>
</tr>
<tr>
<td>W2V2-Transformer *</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td></td>
<td>22.3</td>
<td>34.3</td>
</tr>
<tr>
<td>Chimera Mem-16</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>25.6</td>
<td>35.0</td>
</tr>
<tr>
<td>Chimera</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>27.1*</td>
<td><strong>35.6</strong></td>
</tr>
</tbody>
</table>
Consecutive Decoding for Speech-to-text Translation

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, Lei Li
We design “COnSecutive Transcription and Translation” (COSTT) based on interpreter’s noting behavior to help the model memory.
Motivation of Better Decoding

**Problem 1:** How to give the decoder hints?

**Idea 1:** Introduce a **consecutive decoder** for trans-trans.

**Problem 2:** Long acoustic sequence is challenging for the encoder!

**Idea 2:** Introduce a **compressed encoder** to relief the model memory.
Advantages of COSTT

- Unified training with both transcript and translation text
- Reduced encoder output size with CTC-guided shrinking
- Able to pre-train the decoder with external MT parallel data
End-to-end Speech Translation via Cross-modal Progressive Training

Rong Ye, Mingxuan Wang, Lei Li

• Link: https://arxiv.org/abs/2104.10380
Idea 1: Multi-task Training

Goal: To fully utilize the existing <Speech, Transcript, Translation> supervision.

Decomposed into three sub-tasks with parallel supervision, ST, ASR and MT.
Idea 2: Using large-scale MT data

Comparison of dataset size between ST and MT

<table>
<thead>
<tr>
<th>Dataset</th>
<th>EN-DE</th>
<th>EN-RU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MuST-C  ST dataset</td>
<td>234K</td>
<td>270K</td>
</tr>
<tr>
<td>WMT16 MT dataset</td>
<td>4.6M</td>
<td>2.5M</td>
</tr>
</tbody>
</table>

🤔 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
- We can add more external MT data to train Transformer encoder & decoder

End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]
Supports inputs of two modalities

- Wav2vec2.0\(^1\) as the acoustic encoder
- We add two convolution layers with 2-stride to shrink the length.

---

Language indicator strategy

• We use language indicators to distinguish different tasks.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Source input</th>
<th>Target output</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT</td>
<td>&lt;en&gt; This is a book.</td>
<td>&lt;fr&gt; c'est un livre.</td>
</tr>
<tr>
<td>ASR</td>
<td>&lt;audio&gt; 🎧</td>
<td>&lt;en&gt; This is a book.</td>
</tr>
<tr>
<td>ST</td>
<td>&lt;audio&gt; 🎧</td>
<td>&lt;fr&gt; c'est un livre.</td>
</tr>
</tbody>
</table>
Progressive Multi-task Training

Large-scale MT pre-training

Using external MT $D_{MT-\text{ext}}$

Multi-task Finetune

Using (1) external MT $D_{MT-\text{ext}}$
(2) $D_{ST}$ with <speech, translation>
(3) $D_{ASR}$ with <speech, transcript>

Progressive:
Don’t stop training $D_{MT-\text{ext}}$
XSTNet achieves State-of-the-art Performance

<table>
<thead>
<tr>
<th>Models</th>
<th>External data</th>
<th>Pre-train tasks</th>
<th>En-De</th>
<th>En-Fr</th>
<th>En-Ru</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer ST [13]</td>
<td>×</td>
<td>ASR</td>
<td>22.8</td>
<td>33.3</td>
<td>15.1</td>
<td>23.7</td>
</tr>
<tr>
<td>AFS [28]</td>
<td>×</td>
<td>×</td>
<td>22.4</td>
<td>31.6</td>
<td>14.7</td>
<td>22.9 (-0.8)</td>
</tr>
<tr>
<td>Dual-Decoder Transf. [15]</td>
<td>×</td>
<td>×</td>
<td>23.6</td>
<td>33.5</td>
<td>15.2</td>
<td>24.1 (+0.4)</td>
</tr>
<tr>
<td>STAST [29]</td>
<td>×</td>
<td>×</td>
<td>23.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tang et al. [2]</td>
<td>MT</td>
<td>ASR, MT</td>
<td>24.8</td>
<td>36.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FAT-ST (Big) [6]</td>
<td>ASR, MT, mono-data†</td>
<td>FAT-MLM</td>
<td>25.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>W-Transf.</td>
<td>audio-only*</td>
<td>SSL*</td>
<td>23.6</td>
<td>34.6</td>
<td>14.4</td>
<td>24.2 (+0.5)</td>
</tr>
<tr>
<td>XSTNet-Base</td>
<td>audio-only*</td>
<td>SSL*</td>
<td>25.5</td>
<td>36.0</td>
<td>16.9</td>
<td>26.1 (+2.4)</td>
</tr>
<tr>
<td>XSTNet-Expand</td>
<td>MT, audio-only*</td>
<td>SSL*, MT</td>
<td>27.8</td>
<td>38.0</td>
<td>18.4</td>
<td>27.8 (+4.1)</td>
</tr>
</tbody>
</table>

Table 2: Performance (BLEU) on MuST-C En-De, En-Fr and En-Ru test sets. †: “Mono-data” means audio-only data from Librispeech, Libri-Light, as well as text-only data from Europarl/Wiki Text; *: “Audio-only” data from Librispeech audio data is used in the pre-training of wav2vec2.0-base module, and “SSL” means the self-supervised learning from unlabeled audio data.

**XSTNet-Base**: Achieves the SOTA in the restricted setup

**XSTNet-Expand**: Goes better by using extra MT data

---

End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]
What is “Cascaded-Strong” system?

Strong ASR model + Large-scale MT data

<table>
<thead>
<tr>
<th>Cascaded-Strong</th>
<th>Model</th>
<th>Training data</th>
<th>Performance (En-De)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
<td>W2V2+ Transformer</td>
<td>MuST-C $D_{ASR}$</td>
<td>WER=13.0</td>
</tr>
<tr>
<td>MT</td>
<td>Transformer-base</td>
<td>WMT + MuST-C $D_{MT}$</td>
<td>BLEU=31.7</td>
</tr>
</tbody>
</table>
VolcTransStudio: Video Translation Platform

实时翻译，自动提示 & 交互式修改

Correct-and-Memorize: Learning to translation from interactive revisions [Rongxiang Weng, Hao Zhou, Shujian Huang, Yifan Xia, Lei Li, Jiajun Chen. IJCAI 19]
Summary

- End-to-end Speech-to-Text works!
- Use external ASR, MT data, and audio/text for auxiliary signals
- Model
  - **LUT**: two-stage encoder, additional BERT KD [Dong et al AAAI 2021a]
  - **Chimera**: Shared semantic space encoder with fixed-size memory [Han et al ACL 2021]
  - **COSTT**: consecutive transcription-translation decoder [Dong et al AAAI 2021b]
- Training technique
  - Audio pre-training: Wave2Vec2.0 [Baevski et al 2021]
  - External MT Pre-training
  - **XSTNet**: Progressive multi-task training [Ye et al Interspeech 2021]
Thanks

neural speech translation toolkit
https://github.com/bytedance/neurst

High performance sequence inference
https://github.com/bytedance/lightseq