Introduction

• What is speech translation?
  • Translate speech in source language to text / speech in target language
Introduction

Why/Where do we need speech translation?

• International conferences (e.g., UN, EU)

• Live video translation (e.g., YouTube, streaming)

• Personal translator (e.g., international travels)

  • Google translate (Conversation)
How to do speech translation?
Cascade Speech Translation

• A pipeline of systems
  • Transcribe source speech into source text
  • Translated source text with a text MT model
  • If the output is speech, synthesize speech from target text
Cascade Speech Translation

Waibel et al., 1991; Woszczyna et al., 1993; Vidal, 1997; Wang and Waibel, 1998; Takezawa et al., 1998; Ney, 1999; Bangalore and Riccardi, 2001; Fu-Hua Liu et al., 2003; Schultz et al., 2004; Matusov et al., 2005; Bertoldi and Federico, 2005; Zhang et al., 2005; Pérez et al., 2007; Sperber et al., 2017, 2019; Zhang et al., 2019; Beck et al., 2019; Black et al., 2002; Sumita et al., 2007
Background:
Speech Processing and Recognition

• Speech Processing
  • How to represent speech → feature extraction

• Automatic Speech Recognition (ASR)
  • Transcribe speech to text in one language
  • Seq2seq task, but input and output have the same order
Feature Extractions

- Short-Term Spectrum
  - (Mel-frequency cepstral coefficients) MFCC
- Convert speech samples to sequence of vectors
Feature Extractions

The diagram illustrates the process of feature extractions, starting with the speech signal. The process includes:

1. Pre-emphasis
2. Windowing
3. Discrete Fourier Transform (DFT)
4. Mel filter-bank
5. Logarithmic transformation
6. Inverse Discrete Fourier Transform (IDFT)
7. Computation of MFCC coefficients
8. Calculation of MFCC delta and delta-delta coefficients
9. Extraction of energy features

The output includes 12 MFCC coefficients, 12 delta MFCC coefficients, 12 delta-delta MFCC coefficients, and an additional energy feature.
Automatic Speech Recognition

- Revisit noisy channel model

\[ \hat{Y} = \arg\max P(Y|X) = \arg\max P(X|Y)P(Y) \]

Diagram:
- Language Model
- Acoustic Model
Automatic Speech Recognition

• Acoustic Model
  • Phone recognizer
  • Gaussian mixture model + hidden Markov model
  • Neural-based models
Automatic Speech Recognition

• Acoustic Model
  • Neural-based models

- Fully connect / recurrent layers
- Pooling layers
- Convolutional layers
Automatic Speech Recognition

• Convolutional layers

Today I would like …
Automatic Speech Recognition

• Seq2Seq model
Automatic Speech Recognition

ESPnet: end-to-end speech processing toolkit

ESPnet is an end-to-end speech processing toolkit for speech recognition and end-to-end text-to-speech.

Recipes

For more information on wav2letter++, see or cite this arXiv paper.

To build the old, pre-consolidation version of wav2letter, checkout the wav2letter v0.2 release, which depends on the old Flashlight v0.2 release. The wavletter-lua project can be found on the wavletter-lua branch, accordingly.

Kaldi's code lives at https://github.com/kaldi-asr/kaldi. To checkout (i.e. clone in the git terminology) the most recent changes, you can use this command git clone https://github.com/kaldi-asr/kaldi or follow the github link and click "Download in zip" on the github page (right hand side of the web page).

To browse the model builds that are available (not many), please click on models.

If you have any suggestion of how to improve the site, please contact me.

Contact
 dpovey@gmail.com
 Phone: 425 247 4129
(Daniel Povey)
Cascade Speech Translation

Heute werde ich über Energie und Klima sprechen

Today I am going to talk about energy and climate

Text Machine Translation

Automatic Speech Recognition

Text-to-Speech Synthesis

(a) Speech-to-Text

(b) Speech-to-Speech

Waibel et al., 1991; Woszczyna et al., 1993; Vidal, 1997; Wang and Waibel, 1998; Takezawa et al., 1998; Ney, 1999; Bangalore and Riccardi, 2001; Fu-Hua Liu et al., 2003; Schultz et al., 2004; Matusov et al., 2005; Bertoldi and Federico, 2005; Zhang et al., 2005; Pérez et al., 2007; Sperber et al., 2017, 2019; Zhang et al., 2019; Beck et al., 2019; Black et al., 2002; Sumita et al., 2007
Cascade Speech Translation

• Pros
  • Easy to build (ASR + MT or ASR + MT + TTS)
  • More training data
    • Different data for ASR and MT

• Cons
  • Model size
  • Inference latency
  • Compounding errors
End-to-end Speech Translation

Heute werde ich über Energie und Klima sprechen

(a) Speech-to-Text

End-to-end Speech-to-Text Translation

(b) Speech-to-Speech

End-to-end Speech-to-Speech Translation

Duong et al. (2016); Berard et al. (2016); Weiss et al. (2017); Bansal et al. (2018); Di Gangi et al. (2019b); Pino et al. (2020); Inaguma et al. (2020)
Automatic Speech Recognition

Decoder with attention mechanism

- Recurrent layers
- Pooling layers
- Convolutional layers

Encoder
End-to-End Speech Translation

Decoder with attention mechanism

Encoder

- Recurrent layers
- Pooling layers
- Convolutional layers
End-to-End Speech Translation

• Pros
  • Small model size
  • Lower inference latency
  • No Compounding errors

• Cons
  • Data!
Data Scarcity

- Data is more difficult to collect and annotate
  - Parallel speech to text / speech

<table>
<thead>
<tr>
<th>Tgt</th>
<th>#Talk</th>
<th>#Sent</th>
<th>Hours</th>
<th>src w</th>
<th>tgt w</th>
</tr>
</thead>
<tbody>
<tr>
<td>De</td>
<td>2,093</td>
<td>234K</td>
<td>408</td>
<td>4.3M</td>
<td>4.0M</td>
</tr>
<tr>
<td>Es</td>
<td>2,564</td>
<td>270K</td>
<td>504</td>
<td>5.3M</td>
<td>5.1M</td>
</tr>
<tr>
<td>Fr</td>
<td>2,510</td>
<td>280K</td>
<td>492</td>
<td>5.2M</td>
<td>5.4M</td>
</tr>
<tr>
<td>It</td>
<td>2,374</td>
<td>258K</td>
<td>465</td>
<td>4.9M</td>
<td>4.6M</td>
</tr>
<tr>
<td>Nl</td>
<td>2,267</td>
<td>253K</td>
<td>442</td>
<td>4.7M</td>
<td>4.3M</td>
</tr>
<tr>
<td>Pt</td>
<td>2,050</td>
<td>211K</td>
<td>385</td>
<td>4.0M</td>
<td>3.8M</td>
</tr>
<tr>
<td>Ro</td>
<td>2,216</td>
<td>240K</td>
<td>432</td>
<td>4.6M</td>
<td>4.3M</td>
</tr>
<tr>
<td>Ru</td>
<td>2,498</td>
<td>270K</td>
<td>489</td>
<td>5.1M</td>
<td>4.3M</td>
</tr>
</tbody>
</table>

Speech Translation Dataset

<table>
<thead>
<tr>
<th>#sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT 16 EN-DE</td>
</tr>
<tr>
<td>WMT 14 EN-FR</td>
</tr>
</tbody>
</table>

Machine Translation Dataset

Data Scarcity

- Data augmentation
- Multi-task
  - Share components with other models (ASR / MT)
- Multi-lingual training
- Pretrained components
- Self-learning
  - Train on synthesized data

End-to-End Speech Translation

<table>
<thead>
<tr>
<th>Model</th>
<th>De</th>
<th>Pt</th>
<th>Fr</th>
<th>Es</th>
<th>Ro</th>
<th>Ru</th>
<th>Nl</th>
<th>It</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer + ASR encoder init.  ^</td>
<td>17.30</td>
<td>20.10</td>
<td>26.90</td>
<td>20.80</td>
<td>16.50</td>
<td>10.50</td>
<td>18.80</td>
<td>16.80</td>
</tr>
<tr>
<td><strong>ESPnet-ST (Transformer)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASR encoder/MT decoder init. + SpecAugment</td>
<td>22.91</td>
<td>28.01</td>
<td>32.69</td>
<td>27.96</td>
<td>21.90</td>
<td>15.75</td>
<td>27.43</td>
<td>23.75</td>
</tr>
<tr>
<td><strong>Cascade</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer ASR $\rightarrow$ Transformer MT$^1$</td>
<td>18.5</td>
<td>21.5</td>
<td>27.9</td>
<td>22.5</td>
<td>16.8</td>
<td>11.1</td>
<td>22.2</td>
<td>18.9</td>
</tr>
<tr>
<td><strong>ESPnet-ST</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer ASR $\rightarrow$ Transformer MT</td>
<td>23.65</td>
<td>29.04</td>
<td>33.84</td>
<td>28.68</td>
<td>22.68</td>
<td>16.39</td>
<td>27.91</td>
<td>24.04</td>
</tr>
</tbody>
</table>

End-to-End Speech Translation

• Open source toolkits

**FAIRSEQ S2T: Fast Speech-to-Text Modeling with FAIRSEQ**

Changhan Wang¹, Yun Tang¹, Xutai Ma¹⁺², Anne Wu¹, Dmytro Okhonko¹, Juan Pino¹

¹Facebook AI  
²Johns Hopkins University  
xutai.ma@jhu.edu  
{changhan,yuntang,xutaima,annewu,oxo,juancarabina}@fb.com

**ESPnet-ST: All-in-One Speech Translation Toolkit**

Hirofumi Inaguma¹  Shun Kiyono²  Kevin Duh³  Shigeki Karita⁴  
Nelson Yalta⁵  Tomoki Hayashi⁶⁺⁷  Shinji Watanabe³  
¹ Kyoto University  ² RIKEN AIP  ³ Johns Hopkins University  
⁴ NTT Communication Science Laboratories  ⁵ Waseda University  
⁶ Nagoya University  ⁷ Human Dataware Lab. Co., Ltd.  
inaguma@iap.ist.i.kyoto-u.ac.jp
Simultaneous Speech Translation

• Start the translation before read all the input speech

I am going to talk today about energy and climate.

Heute spreche ich zu Ihnen über Energie und Klima.
Simultaneous Speech Translation

Input: I am going to talk today about energy and climate.

Output: Heute spreche ich zu Ihnen über Energie und Klima.
Simultaneous Translation Policies

- **Reinforcement learning** (Gu et al. 2017; Luo et al. 2017; Lawson et al. 2018)
  - Less stable learning process.

- **Fixed policy** (Cho and Esipova 2016; Ma et al. 2019a)
  - Weaker performance, for instance Wait-K (Ma et al. 2019a).

- **Monotonic attention** (Raffel et al., 2017; Arivazha-gan et al., 2019; Ma et al., 2020)
  - The State of the art for the task.
Quality-Latency Trade-off

The lagging behind an oracle/perfect system
Speech-to-Speech Translation

- Source speech $\rightarrow$ target speech
- Cascade
  - ST + Text-to-speech (TTS) system
- End-to-end (direct)
  - Directly generate target spectrogram
  - Preserve prosody, emphasis, emotion
  - Suffers more from data scarcity
Direct Speech-to-Speech Translation

Source Speech signals → Feature Extractor → Source Speech Features

Target Speech signals ← Vocoder ← Source Speech Spectrogram

S2ST Model
Direct S2ST with Sequence-to-Sequence Model

- Speech encoder from ASR & ST
- Spectrogram decoder from TTS
- Multi-task learning
- Examples

Y. Jia et al., "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model." Proc. Interspeech 2019
Conclusion

• Speech translation is a new and challenging task

• End-to-end approach shows strength when overcome the data scarcity issue.

• More challenges: low latency, speech-to-speech, etc.