Sequence-to-Sequence Models Can Directly Translate Foreign Speech

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, Zhifeng Chen
End-to-end training for speech translation

- **Task:** Spanish speech to English text translation
  - Typically train specialized translation model on ASR output lattice, or integrate ASR and translation decoding using e.g. stochastic FST

- **Why end-to-end?**
  - Directly optimize for desired output, avoid compounding errors
    - e.g. difficult for text translation system to recover from gross misrecognition
  - Single decoding step -> low latency inference
  - Less training data required -- don't need both transcript *and* translations
    - (might not be an advantage)

- **Use sequence-to-sequence neural network model**
  - Flexible framework, easily admits multi-task training
  - Previous work
    - [Bérard et al, 2016] trained "Listen and Translate" seq2seq model on *synthetic speech*
    - [Duong et al, 2016] seq2seq model to *align* speech with translation
Sequence-to-sequence / Encoder-decoder with attention

- Recurrent neural net that maps between arbitrary length sequences [Bahdanau et al, 2015]
  - e.g. "Listen, Attend and Spell" [Chan et al, 2016] and [Chorowski et al, 2015]
    sequence of spectrogram frames -> sequence of characters
Stacked (bidirectional) RNN computes *latent representation* of input sequence

- Following [Zhang et al, 2017], include convolutional layers to downsample sequence in time
Decoder RNN

- Autoregressive next-step prediction -- outputs one character at a time
- Conditioned on entire encoded input sequence via attention context vector
Attention

- For each output token, generates a **context vector** from encoder latent representation
- Computes an alignment between input and output sequences
  - \( \text{Prob}(h_i | y_{1..k}) \)
Seq2seq ASR: Architecture details

- **Input**: 80 channel log mel filterbank features
  - + deltas and accelerations

- **Encoder** follows [Zhang et al, 2017]
  - 2 stacked 3x3 convolution layers, strided to downsample in time by a total factor of 4
  - 1 convolutional LSTM layer
  - 3 stacked bidirectional LSTM layers with 512 cells
  - batch normalization

- **Additive attention** [Bahdanau et al, 2015]

- **Decoder**
  - 4 stacked unidirectional LSTM layers
    - >= 2 layers improve performance, especially for speech translation
  - skip connections pass attention context to each decoder layer

- **Regularization**: Gaussian weight noise and L2 weight decay
Compare three approaches:

1. **ASR -> NMT cascade**
   - train independent Spanish ASR, and text neural machine translation models
   - pass top ASR hypothesis through NMT

2. **End-to-end ST**
   - train LAS model to directly predict English text from Spanish audio
   - identical architectures for Spanish and Spanish-English ST

3. **Multi-task ST / ASR**
   - shared encoder
   - 2 independent decoders with different attention networks
     - each emits text in a different language
Seq2seq Speech Translation (ST): End-to-end

Compare three approaches:

1. **ASR -> NMT cascade**
   - train independent Spanish ASR, and text neural machine translation models
   - pass top ASR hypothesis through NMT

2. **End-to-end ST**
   - train LAS model to directly predict *English text* from Spanish audio
   - identical architectures for Spanish ASR and Spanish-English ST

3. **Multi-task ST / ASR**
   - *shared* encoder
   - 2 independent decoders with *different* attention networks
     - each emits text in a different language
Compare three approaches:

1. **ASR -> NMT cascade**
   - Train independent Spanish ASR, and text neural machine translation models
   - Pass top ASR hypothesis through NMT

2. **End-to-end ST**
   - Train LAS model to directly predict English text from Spanish audio
   - Identical architectures for Spanish ASR and Spanish-English ST

3. **Multi-task ST / ASR**
   - *shared* encoder
   - 2 independent decoders with *different* attention networks
   - Each emits text in a different language
Seq2seq Speech Translation: Attention

- recognition attention very confident
- translation attention smoothed out across many spectrogram frames for each output character
  - ambiguous mapping between Spanish speech acoustics and English text
Seq2seq Speech Translation: Attention

- speech recognition attention is mostly monotonic
- translation attention reorders input: same frames attended to for "vive aqui" and "living here"
Experiments: Fisher/Callhome Spanish-English data

- Transcribed Spanish telephone conversations from LDC
  - Fisher: conversations between strangers
  - Callhome: conversations between friends and family. more informal and challenging

- Crowdsourced English translations of Spanish transcripts from [Post et al, 2013]

- Train on 140k Fisher utterances (160 hours)

- Tune using Fisher/dev

- Evaluate on held out Fisher/test set and Callhome
## Experiments: Baseline models

<table>
<thead>
<tr>
<th></th>
<th>dev</th>
<th>Fisher dev2</th>
<th>test</th>
<th>Callhome devtest</th>
<th>evltest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours$^3$</td>
<td>25.7</td>
<td>25.1</td>
<td>23.2</td>
<td>44.5</td>
<td>45.3</td>
</tr>
<tr>
<td>Post et al. [19]</td>
<td>41.3</td>
<td>40.0</td>
<td>36.5</td>
<td>64.7</td>
<td>65.3</td>
</tr>
<tr>
<td>Kumar et al. [21]</td>
<td>29.8</td>
<td>29.8</td>
<td>25.3</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

- **WER** on Spanish ASR
  - seq2seq model outperforms classical GMM-HMM [19] and DNN-HMM [21] baselines

<table>
<thead>
<tr>
<th></th>
<th>dev</th>
<th>Fisher dev2</th>
<th>test</th>
<th>Callhome devtest</th>
<th>evltest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>58.7</td>
<td>59.9</td>
<td>57.9</td>
<td>28.2</td>
<td>27.9</td>
</tr>
<tr>
<td>Post et al. [19]</td>
<td>–</td>
<td>–</td>
<td>58.7</td>
<td>–</td>
<td>27.8</td>
</tr>
<tr>
<td>Kumar et al. [21]</td>
<td>–</td>
<td>65.4</td>
<td>62.9</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

- **BLEU score** on Spanish-to-English text translation
  - seq2seq NMT (following [Wu et al, 2016]) slightly underperforms phrase-based SMT baselines
Experiments: End-to-end speech translation

<table>
<thead>
<tr>
<th>Model</th>
<th>dev</th>
<th>Fisher dev2</th>
<th>test</th>
<th>Callhome devtest</th>
<th>callhome evltest</th>
</tr>
</thead>
<tbody>
<tr>
<td>End-to-end ST$^3$</td>
<td>46.5</td>
<td>47.3</td>
<td>47.3</td>
<td>16.4</td>
<td>16.6</td>
</tr>
<tr>
<td>Multi-task ST / ASR$^3$</td>
<td>48.3</td>
<td>49.1</td>
<td>48.7</td>
<td>16.8</td>
<td>17.4</td>
</tr>
<tr>
<td>ASR $\rightarrow$ NMT cascade$^3$</td>
<td>45.1</td>
<td>46.1</td>
<td>45.5</td>
<td>16.2</td>
<td>16.6</td>
</tr>
<tr>
<td>Post et al. [19]</td>
<td>–</td>
<td>35.4</td>
<td>–</td>
<td>–</td>
<td>11.7</td>
</tr>
<tr>
<td>Kumar et al. [21]</td>
<td>–</td>
<td>40.1</td>
<td>40.4</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

- BLEU score (higher is better)
- Multi-task $>$ End-to-end ST $>$ Cascade $>>$ non-seq2seq baselines
- ASGD training with 10 replicas (16 for multitask)
  - ASR model converges after 4 days
  - ST and multi-task models continue to improve for 2 weeks
Example output: compounding errors

**ASR**
ref: "sí a mime gusta mucho bailar merengue y salsa también"

hyp: "sea me gusta mucho bailar merengue y sabes también"
hyp: "sea me gusta mucho bailar medio inglés"
hyp: "o sea me gusta mucho bailar merengue y sabes también"
hyp: "sea me gusta mucho bailar medio inglés sabes también"
hyp: "sea me gusta mucho bailar merengue"
hyp: "o sea me gusta mucho bailar medio inglés"
hyp: "sea no gusta mucho bailar medio inglés"
hyp: "o sea me gusta mucho bailar medio inglés sabes también"

**End-to-end ST**
ref: "yes i do enjoy dancing merengue and salsa music too"

hyp: "i really like to dance merengue and salsa also"
hyp: "i like to dance merengue and salsa also"
hyp: "i don't like to dance merengue and salsa also"
hyp: "i really like to dance merengue and salsa and also"
hyp: "i like to dance merengue and salsa and also"
hyp: "i really like to dance merengue and salsa"
hyp: "i like to dance merengue and salsa"
hyp: "i don't like to dance merengue and salsa and also"

**Cascade: ASR top hypothesis -> NMT**
hyp: "i really like to dance merengue and you know also"

- ASR consistently mis-recognizes "merengue y salsa" as "merengue y sabes" or "medio inglés"
- NMT has no way to recover
Conclusions

● **Proof of concept end-to-end model for conversational speech translation**
  ○ without performing intermediate speech recognition,
    nor requiring any supervision from the source language transcripts*
  ○ without explicitly training or tuning separate language model or text translation model
  ○ no need to optimize model combination

● **Identical model architecture and beam search decoding algorithm**
can be used for both speech recognition and translation
  ○ it turns out that sequence-to-sequence models are quite powerful

● **Can further improve performance by multi-task training ASR and ST models**
  ○ regularization effect of encouraging encoder to learn a representation suitable for both tasks
References


Extra slides
End-to-end model: tuning

- Performance improves with deeper decoder

<table>
<thead>
<tr>
<th>Num decoder layers $D$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>43.8</td>
<td>45.1</td>
<td>45.2</td>
<td>45.5</td>
<td>45.3</td>
</tr>
</tbody>
</table>

- Best speech translation performance in multitasked model when full encoder is shared across both tasks

<table>
<thead>
<tr>
<th>Num shared encoder LSTM layers</th>
<th>3 (all)</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>46.2</td>
<td>45.1</td>
<td>45.3</td>
<td>44.2</td>
</tr>
</tbody>
</table>
Seq2seq Speech Translation: Example attention

- translation model attends to the beginning of input (i.e. silence) for the last few letters in each word
  - already made a decision about word to emit, just acts a language model to spell it out.