Pre-training on high-resource speech recognition improves low-resource speech-to-text translation

Sameer Bansal
Herman Kamper
Karen Livescu
Adam Lopez
Sharon Goldwater
Current systems

Spanish Audio: 

English text: ?
Current systems

Spanish Audio:  

Spanish text:  *ola mi nombre es hodor*

English text:  ?

Automatic Speech Recognition
Current systems

Spanish Audio:

Spanish text: ola mi nombre es hodor

English text: hi my name is hodor

Automatic Speech Recognition

Machine Translation
~100 languages supported by Google Translate ...
Unwritten languages

Mboshi:  

Bantu language, Republic of Congo, ~160K speakers

~3000 languages with no writing system

Mboshi text:  not available  

Automatic Speech Recognition
Unwritten languages

Mboshi: ~3000 languages with no writing system

Efforts to collect speech and translations using mobile apps

Haiti Earthquake, 2010

Survivors sent text messages to helpline

- Moun kwense nan
- Sakre Kè nan
- Pòtoprens

- People trapped in Sacred Heart Church, PauP

- International rescue teams face language barrier
- No automated tools available
- Volunteers from global Haitian diaspora help create parallel text corpora in short time

[Munro 2010]
Are we better prepared in 2019?

Moun kwense nan Sakre
Kè nan Pòtoprens

People trapped in Sacred Heart Church, PauP

Voice messages
Can we build a speech-to-text translation (ST) system?

… given as training data:

- Tens of hours of speech paired with text translations
- No source text available
Neural models ...

Spanish Audio:

Sequence-to-Sequence

Weiss et al. (2017)

English text:  hi my name is hodor

Directly translate speech
Spanish speech to English text

- telephone speech (unscripted)
- realistic noise conditions
- multiple speakers and dialects
- crowdsourced English text translations

Closer to real-world conditions
Spanish speech to English text

Good performance if trained on 100+ hours

Weiss et al.

*for comparison text-to-text = 58
Poor performance in
low-resource settings

*for comparison
text-to-text = 58
Goal: to improve translation performance
Goal: to improve translation performance

... without labeling more low-resource speech
100s of hours of **monolingual** speech paired with text available

... typically used to train ASR systems

Key idea: leverage monolingual data from a different high-resource language
100s of hours of **monolingual** speech paired with text available

... typically used to train ASR systems

(English Audio)  
(English text)  
(French Audio)  
(French text)

**Spanish** Audio  
Sequence-to-Sequence  
**English** text

~20 hours of **Spanish-English**
100s of hours of monolingual speech paired with text available

... typically used to train ASR systems

Weiss et al. 2017
Anastasopoulos and Chiang 2018
Bérard et al. 2018
Sperber et al. 2019

~20 hours of Spanish-English
100s of hours of **monolingual** speech paired with text available

... typically used to train ASR systems

[Diagram showing English text (English Audio) and French text (French Audio)]

**Spanish** Audio → **Sequence-to-Sequence** → **English** text

≈20 hours of **Spanish-English**
Why Spanish-English?
Why Spanish-English?

simulate low-resource settings and test our method
Why Spanish-English?

simulate low-resource settings and test our method

Later: results on truly low-resource language --- Mboshi to French
Method

Audio

Encoder

Attention

Decoder

text

Same model architecture for ASR and ST

*randomly initialized parameters
Pretrain on high-resource

English audio

Encoder

Attention

Decoder

English text

300 hours of English audio and text

*train until convergence
Fine-tune on low-resource

20 hours Spanish-English

Diagram:

- English audio
  - Encoder
  - Attention
  - Decoder
  - English text

- Spanish audio
  - Encoder
  - Attention
  - Decoder
  - English text

Transfer from English ASR
Fine-tune on low-resource

20 hours Spanish-English

*train until convergence

Spanish audio

Encoder

Attention

Decoder

English text
Will this work?
Spanish-English BLEU scores

*for comparison
Weiss et al. = 47.3
Spanish-English BLEU scores

*for comparison
Weiss et al. = 47.3
Spanish-English BLEU scores

*for comparison
Weiss et al. = 47.3

pretraining

baseline

+9 BLEU

31
Spanish-English BLEU scores

- Better performance with half the data

*For comparison
Weiss et al. = 47.3
Further analysis

20 hours Spanish-English

*for comparison
Weiss et al. = 47.3
Faster training time

**Graph**

- **Y-axis**: BLEU
- **X-axis**: training epochs

**Lines**
- **Pretraining**: solid green line
- **Baseline**: dashed purple line

**Legend**
- *pretraining*
- *baseline*
Faster training time

- potentially useful in time critical scenarios

2 hours

~20 hours

pretraining

baseline
Spanish to English, $N = 20$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>10.8</td>
</tr>
<tr>
<td>+English ASR</td>
<td>19.9</td>
</tr>
</tbody>
</table>

Ablation: model parameters

English

- Encoder
- Attention
- Decoder

English text

Spanish

- Encoder
- Attention
- Decoder

English text
Ablation: model parameters

Spanish to English, $N = 20$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>10.8</td>
</tr>
<tr>
<td>+English ASR</td>
<td>19.9</td>
</tr>
<tr>
<td>+English ASR: decoder</td>
<td>10.5</td>
</tr>
</tbody>
</table>
Ablation: model parameters

Spanish to English, $N = 20$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>10.8</td>
</tr>
<tr>
<td>+English ASR</td>
<td>19.9</td>
</tr>
<tr>
<td>+English ASR: decoder</td>
<td>10.5</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>16.6</td>
</tr>
</tbody>
</table>

English

Spanish

random
Ablation: model parameters

Spanish to English, $N = 20$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>10.8</td>
</tr>
<tr>
<td>+English ASR</td>
<td>19.9</td>
</tr>
<tr>
<td>+English ASR: decoder</td>
<td>10.5</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>16.6</td>
</tr>
</tbody>
</table>

... transferring encoder only parameters works well!
### Ablation: model parameters

#### Spanish to English, \( N = 20 \) hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>10.8</td>
</tr>
<tr>
<td>+English ASR</td>
<td>19.9</td>
</tr>
<tr>
<td>+English ASR: decoder</td>
<td>10.5</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>16.6</td>
</tr>
</tbody>
</table>

… can pretrain on a language different from both source and target in ST pair
Pretraining on French

Spanish to English, $N = 20$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>10.8</td>
</tr>
<tr>
<td>+English ASR</td>
<td>19.9</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>16.6</td>
</tr>
<tr>
<td>+French ASR: encoder</td>
<td>?</td>
</tr>
</tbody>
</table>

*only 20 hours of French ASR
Pretraining on French

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>10.8</td>
</tr>
<tr>
<td>+English ASR</td>
<td>19.9</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>16.6</td>
</tr>
<tr>
<td>+French ASR: encoder</td>
<td><strong>12.5</strong></td>
</tr>
</tbody>
</table>

Spanish to English, $N = 20$ hours

French ASR helps Spanish-English ST
Takeaways

- Pretraining on a different language helps
- Transfer all model parameters for best gains
- Encoder parameters account for most of these

... useful when target vocabulary is different
… Mboshi-French ST
Mboshi-French ST

- ST data by Godard et al. 2018
  - ~4 hours of speech, paired with French translations

- Mboshi
  - Bantu language, Republic of Congo
  - Unwritten
  - ~160K speakers
Mboshi-French: Results

Mboshi to French, $N = 4$ hours

| BLEU  | baseline | ? |

Mboshi

Encoder

Attention

Decoder

French text
Mboshi-French: Results

Mboshi to French, $N = 4$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.5</td>
</tr>
</tbody>
</table>

*outperformed by a naive baseline

 Mboshi

 Encoder

 Attention

 Decoder

 French text
### Pretraining on French ASR

<table>
<thead>
<tr>
<th>Mboshi to French, $N = 4$ hours</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.5</td>
</tr>
<tr>
<td>+French ASR: all</td>
<td>?</td>
</tr>
</tbody>
</table>

- **Transfer all parameters**
Pretraining on French ASR

Mboshi to French, $N = 4$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.5</td>
</tr>
<tr>
<td>+French ASR: all</td>
<td>5.9</td>
</tr>
</tbody>
</table>

French ASR helps Mboshi-French ST
Pretraining on French ASR

Mboshi to French, $N = 4$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.5</td>
</tr>
<tr>
<td>+French ASR: all</td>
<td>5.9</td>
</tr>
</tbody>
</table>

French ASR helps Mboshi-French ST
Pretraining on English ASR

<table>
<thead>
<tr>
<th>Mboshi to French, $N = 4$ hours</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.5</td>
</tr>
<tr>
<td>+French ASR: all</td>
<td>5.9</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>?</td>
</tr>
</tbody>
</table>

using encoder trained on a lot more data
Pretraining on English ASR

Mboshi to French, $N = 4$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.5</td>
</tr>
<tr>
<td>+French ASR: all</td>
<td>5.9</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>5.3</td>
</tr>
</tbody>
</table>

English ASR helps Mboshi-French ST
Pretraining on French ASR: can transfer all parameters

… but only 20 hours of data

Pretraining on English ASR: trained on a lot more data (300 hours)

… but can only transfer encoder parameters
Pretraining on French ASR: can transfer all parameters

... but only 20 hours of data

Pretraining on English ASR: trained on a lot more data (300 hours)

... but can only transfer encoder parameters

... combine both?
Pretraining on French and English ASR

French
- Encoder
- Attention
- Decoder
  French text
  20 hours

English
- Encoder
- Attention
- Decoder
  English text
  300 hours
Pretraining on French and English ASR

French
Encoder
Attention
Decoder
French text

20 hours

Mboshi
Encoder
Attention
Decoder
French text

4 hours

English
Encoder
Attention
Decoder
English text

300 hours
Pretraining on French and English ASR

- **French**
  - Encoder
  - Attention
  - Decoder
  - French text
  - 20 hours

- **Mboshi**
  - Encoder
  - Attention
  - Decoder
  - French text
  - 4 hours

- **English**
  - Encoder
  - Attention
  - Decoder
  - English text
  - 300 hours
### Pretraining on English ASR

Mboshi to French, $N = 4$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.5</td>
</tr>
<tr>
<td>+French ASR: all</td>
<td>5.9</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>5.3</td>
</tr>
<tr>
<td>+English ASR: encoder +French ASR: remaining</td>
<td>?</td>
</tr>
</tbody>
</table>

---

![Diagram showing the flow of Mboshi from English ASR and French ASR to French text via Encoder, Attention, and Decoder.]
**Pretraining on English ASR**

Mboshi to French, $N = 4$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.5</td>
</tr>
<tr>
<td>+French ASR: all</td>
<td>5.9</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>5.3</td>
</tr>
<tr>
<td>+English ASR: encoder +French ASR: remaining</td>
<td>7.1</td>
</tr>
</tbody>
</table>

Combining gives the best gains
Pretraining on English ASR

Mboshi to French, $N = 4$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.5</td>
</tr>
<tr>
<td>+French ASR: all</td>
<td>5.9</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>5.3</td>
</tr>
<tr>
<td>+English ASR: encoder</td>
<td>7.1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BLEU score is still low … but above naive baseline
Conclusions

● Pretraining on high-resource ASR improves low-resource ST
● Potentially useful for endangered and/or unwritten languages
● Bootstrap ST in time-critical scenarios
● Future work: experiments on more languages, multilingual training with joint vocabulary
Thanks

• Anonymous reviewers, Edinburgh NLP members
• Source code available at: https://github.com/0xSameer/ast

I am looking for full-time positions starting November 2019!

• 4th June, 3:30-5 pm - “Fluent Translations from Disfluent Speech in End-to-End Speech Translation”, Salesky et al.
• 5th June, 10:30-10:48 am - “Neural Machine Translation of Text from Non-Native Speakers”, Anastasopoulos et al.
Backup
# Mboshi-French naive baseline

<table>
<thead>
<tr>
<th>model</th>
<th>pretrain</th>
<th>BLEU</th>
<th>Pr.</th>
<th>Rec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr-top-8w</td>
<td>–</td>
<td>0</td>
<td>23.5</td>
<td>22.2</td>
</tr>
<tr>
<td>fr-top-10w</td>
<td>–</td>
<td>0</td>
<td>20.6</td>
<td>24.5</td>
</tr>
<tr>
<td>en-300h</td>
<td>–</td>
<td>0</td>
<td>0.2</td>
<td>5.7</td>
</tr>
<tr>
<td>fr-20h</td>
<td>–</td>
<td>0</td>
<td>4.1</td>
<td>3.2</td>
</tr>
</tbody>
</table>

mb-fr-4h

<table>
<thead>
<tr>
<th>model</th>
<th>pretrain</th>
<th>BLEU</th>
<th>Pr.</th>
<th>Rec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mb-fr-4h</td>
<td>fr-20h</td>
<td>5.9</td>
<td>23.6</td>
<td>20.9</td>
</tr>
<tr>
<td></td>
<td>en-300h</td>
<td>5.3</td>
<td>23.5</td>
<td>22.6</td>
</tr>
<tr>
<td></td>
<td>en + fr</td>
<td>7.1</td>
<td>26.7</td>
<td>23.1</td>
</tr>
</tbody>
</table>
Why does pretraining help?

- Speaker invariance
  - ASR data contains audio from 100s of speakers
- Learning to factor out background noise (?)

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Baseline</th>
<th>+English ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 speakers</td>
<td>7.2</td>
<td>17.5 (+143 %)</td>
</tr>
<tr>
<td>136 speakers</td>
<td>10.8 (+ 50%)</td>
<td>19.9 (+14%)</td>
</tr>
</tbody>
</table>
**Spanish-English ST**

<table>
<thead>
<tr>
<th>$N$ hrs</th>
<th>2.5h</th>
<th>5h</th>
<th>10h</th>
<th>20h</th>
<th>50h</th>
<th>160h Weiss</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>2.1</td>
<td>1.8</td>
<td>2.1</td>
<td>10.8</td>
<td>22.7</td>
<td>47.3</td>
</tr>
<tr>
<td>+ASR</td>
<td>5.7</td>
<td>9.1</td>
<td>14.5</td>
<td>20.2</td>
<td>28.3</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>+3.6</td>
<td>+7.3</td>
<td>+12.4</td>
<td>+9.4</td>
<td>+5.5</td>
<td></td>
</tr>
</tbody>
</table>

*results on Fisher test set ...
## Spanish-English ST

Spanish to English, $N = 20$ hours

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>10.8</td>
</tr>
<tr>
<td>+En ASR: 300h</td>
<td>16.6</td>
</tr>
<tr>
<td>+Fr ASR: 20h</td>
<td><strong>12.5</strong></td>
</tr>
<tr>
<td>+En ASR: 20h</td>
<td>13.2</td>
</tr>
</tbody>
</table>

… French ASR helps improve Spanish-English ST
Neural model

MFCCs
150 x 13

75 x 128

37 x 512

37 x 512

CNN 1

CNN 2

bi-LSTM 1

bi-LSTM 2

bi-LSTM 3

Attention

FF-Softmax

LSTM 1

LSTM 2

LSTM 3

Embedding

i live in br_ __ on_ __ x EOS

i live in bronx

1.5 s
Neural model

Encoder

Decoder

predicted text

FF-Softmax

RNN

Embedding

prediction history

Attention

CNN

RNN

MFCCs
100s of hours of **monolingual** speech paired with text available

... typically used to train ASR systems

Gülçehre et al., 2015
Toshniwal et al., 2018

Spanish Audio  Sequence-to-Sequence  English text

~20 hours of Spanish-English