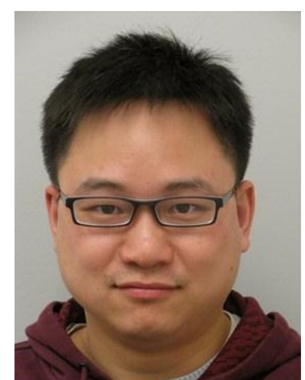
# Sequence-to-Sequence Models Can Directly Translate Foreign Speech

Ron J. Weiss, **Jan Chorowski**, Navdeep Jaitly, Yonghui Wu, Zhifeng Chen



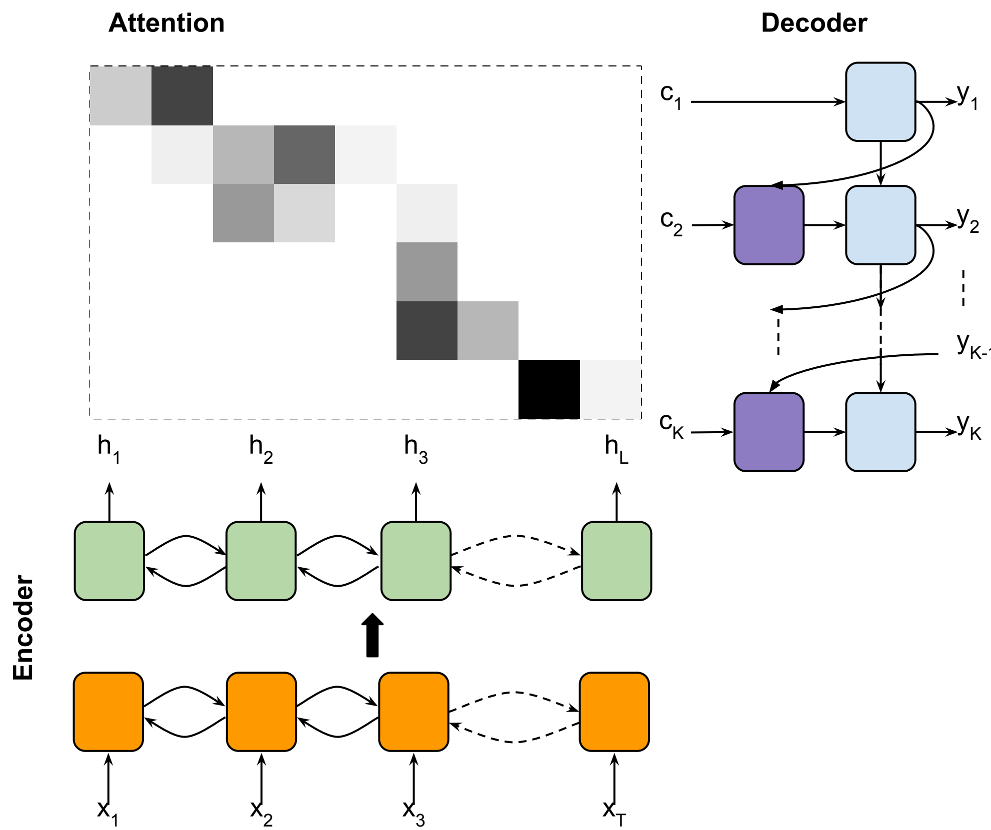
# End-to-end training for speech translation

- 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)

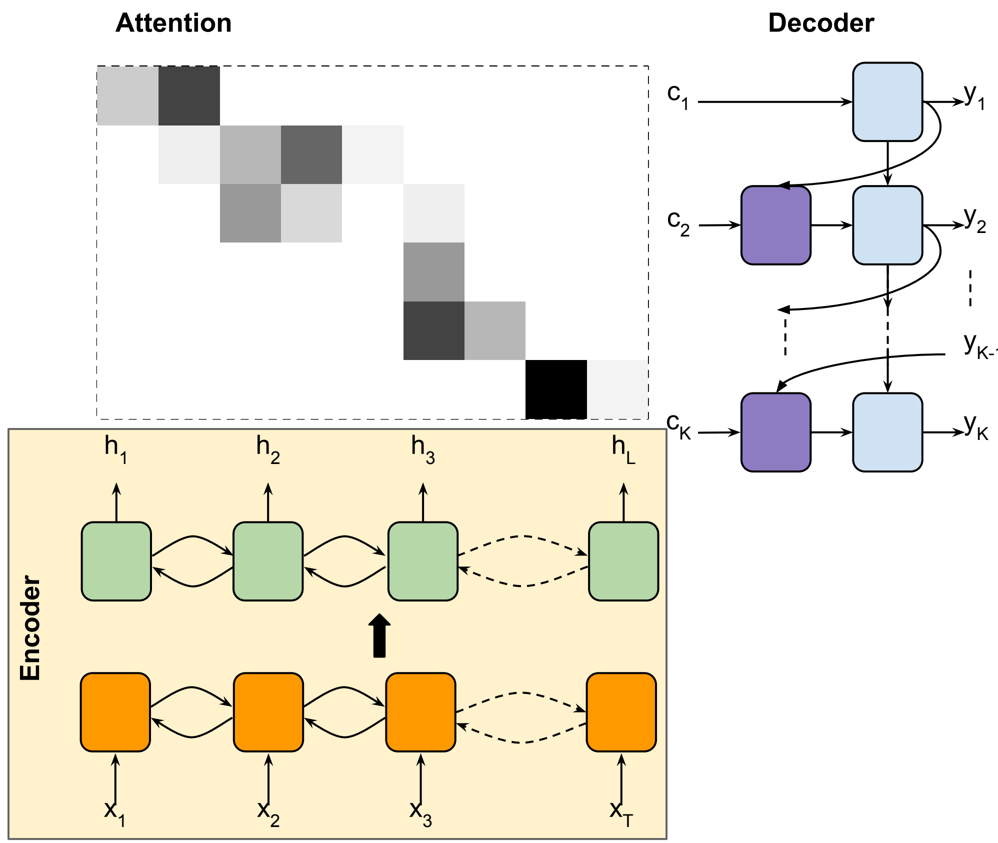
- Use sequence-to-sequence neural network model
  - Flexible framework, easily admits multi-task training
  - Previous work
    - [Bérard et al, 2016] trained "Listen and Translate" seq2seq model on *synthetic speech*
    - [Duong et al, 2016] seq2seq model to *align* speech with translation

# Sequence-to-sequence / Encoder-decoder with attention



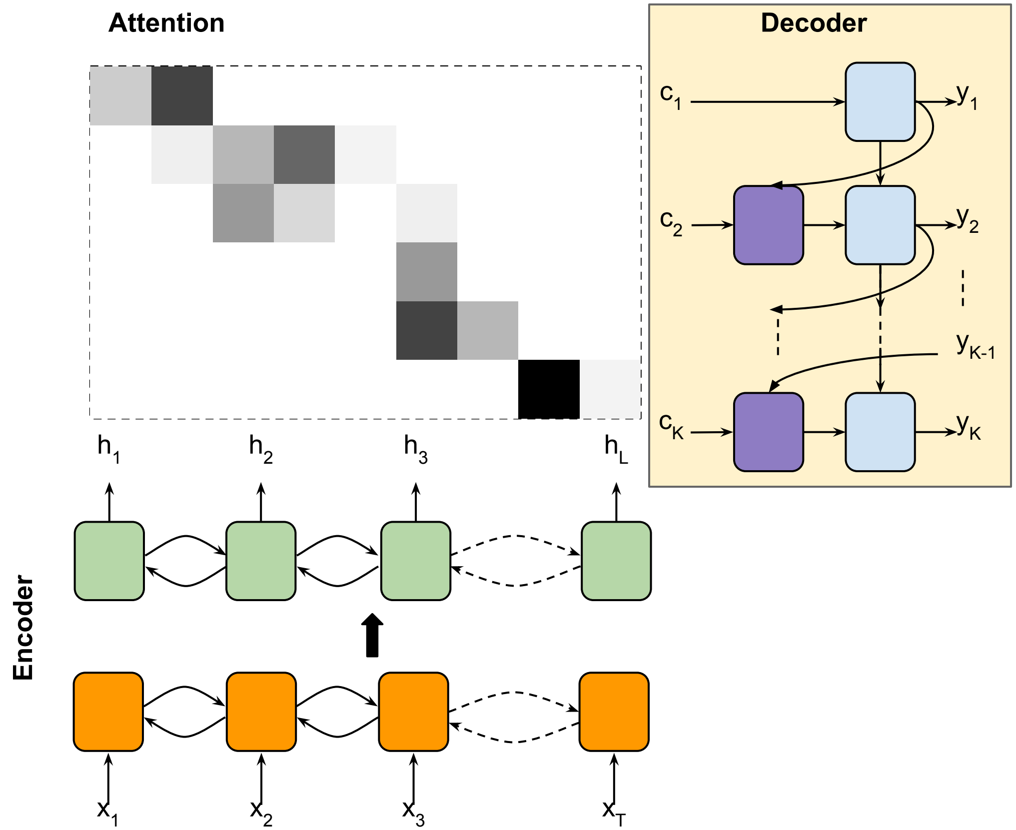

- Recurrent neural net that maps between arbitrary length sequences [Bahdanau et al, 2015]
  - e.g. "Listen, Attend and Spell" [Chan et al, 2016] and [Chorowski et al, 2015] sequence of spectrogram frames -> sequence of characters

# Encoder RNN



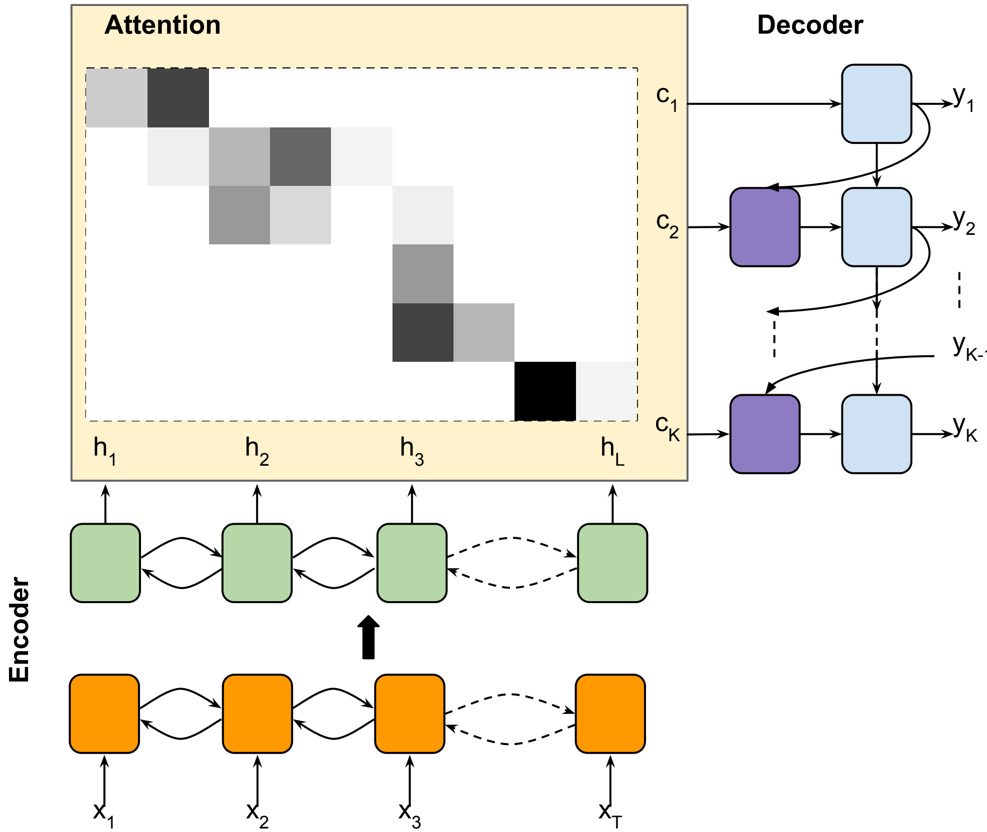

- Stacked (bidirectional) RNN computes *latent representation* of input sequence
  - Following [Zhang et al, 2017], include convolutional layers to downsample sequence in time

# Decoder RNN




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

# Attention

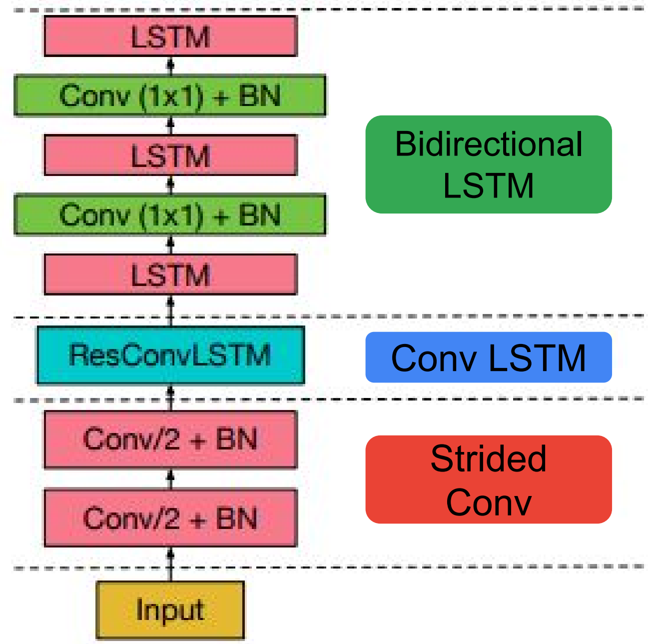



- For each output token, generates a *context vector* from encoder latent representation
- Computes an alignment between input and output sequences
  - $\mathrm{Prob}(h_i \mid y_{1..k})$

Google

# Seq2seq ASR: Architecture details

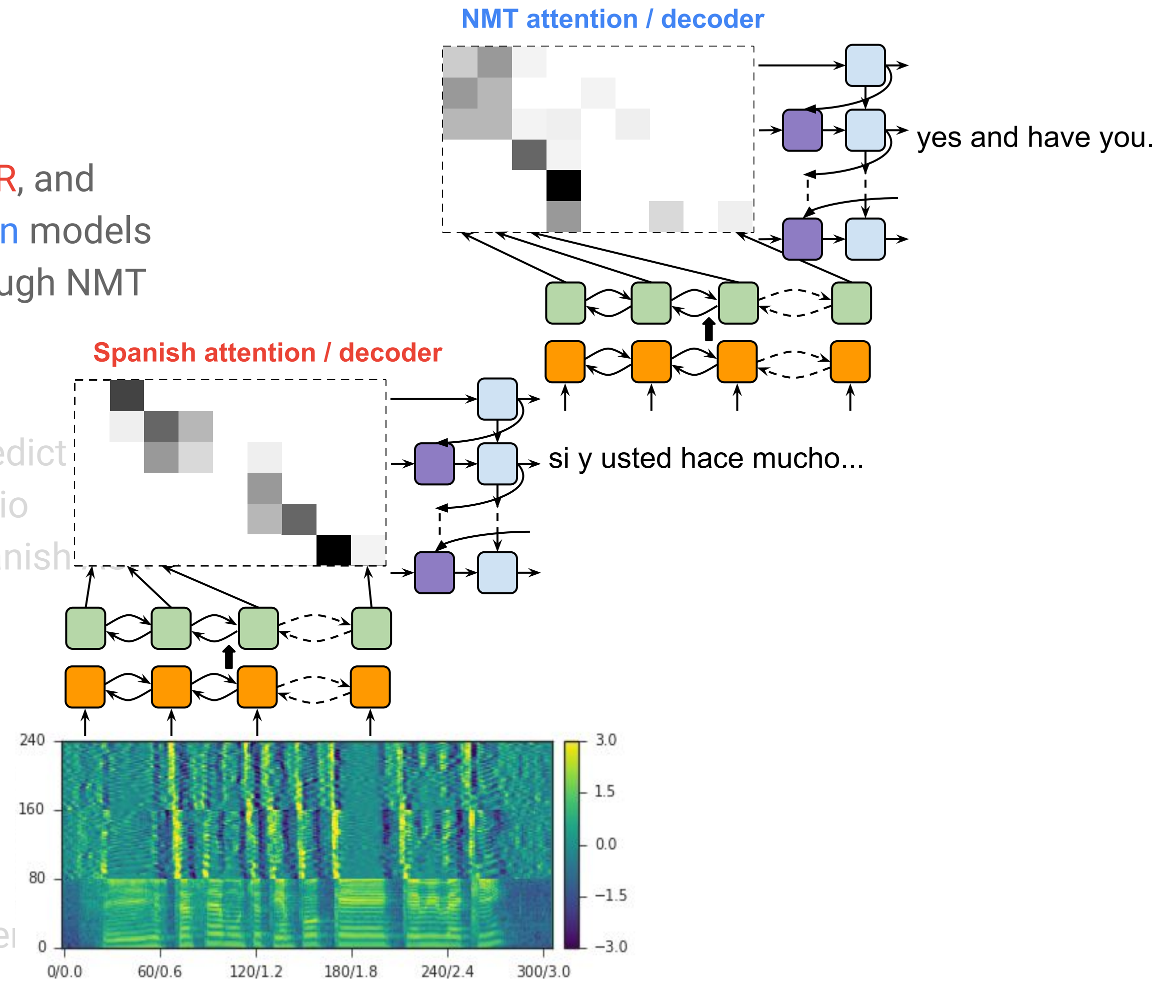
- Input: 80 channel log mel filterbank features
  - + deltas and accelerations
- Encoder follows [Zhang et al, 2017]
  - 2 stacked 3x3 convolution layers,
    strided to downsample in time by a total factor of 4
  - 1 convolutional LSTM layer
  - 3 stacked bidirectional LSTM layers with 512 cells
  - 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



# Seq2seq Speech Translation (ST): Cascade

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 differe



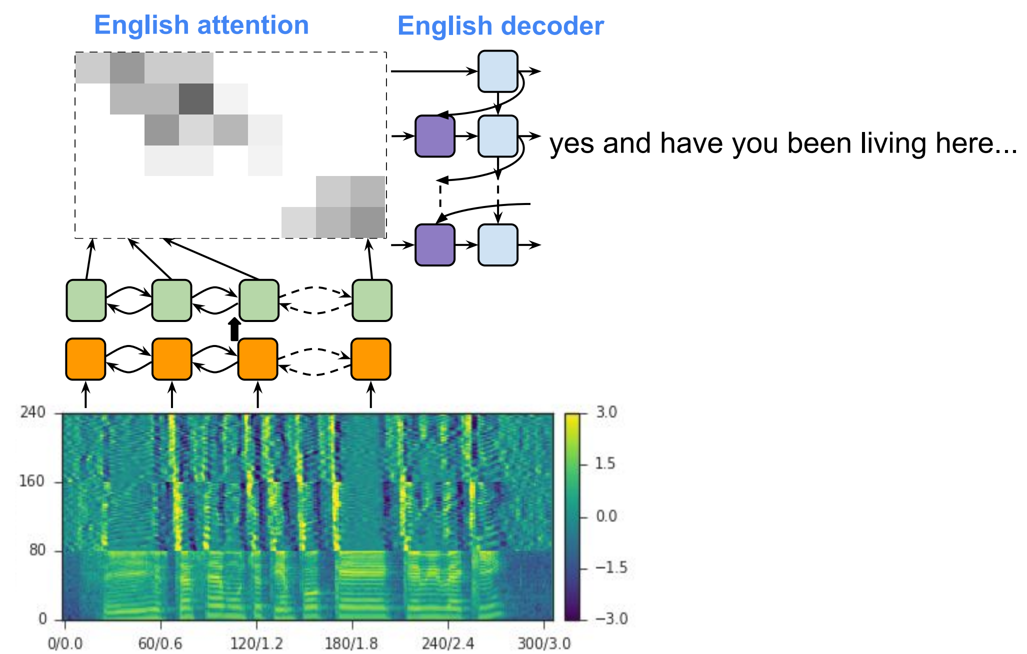

Google

# 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

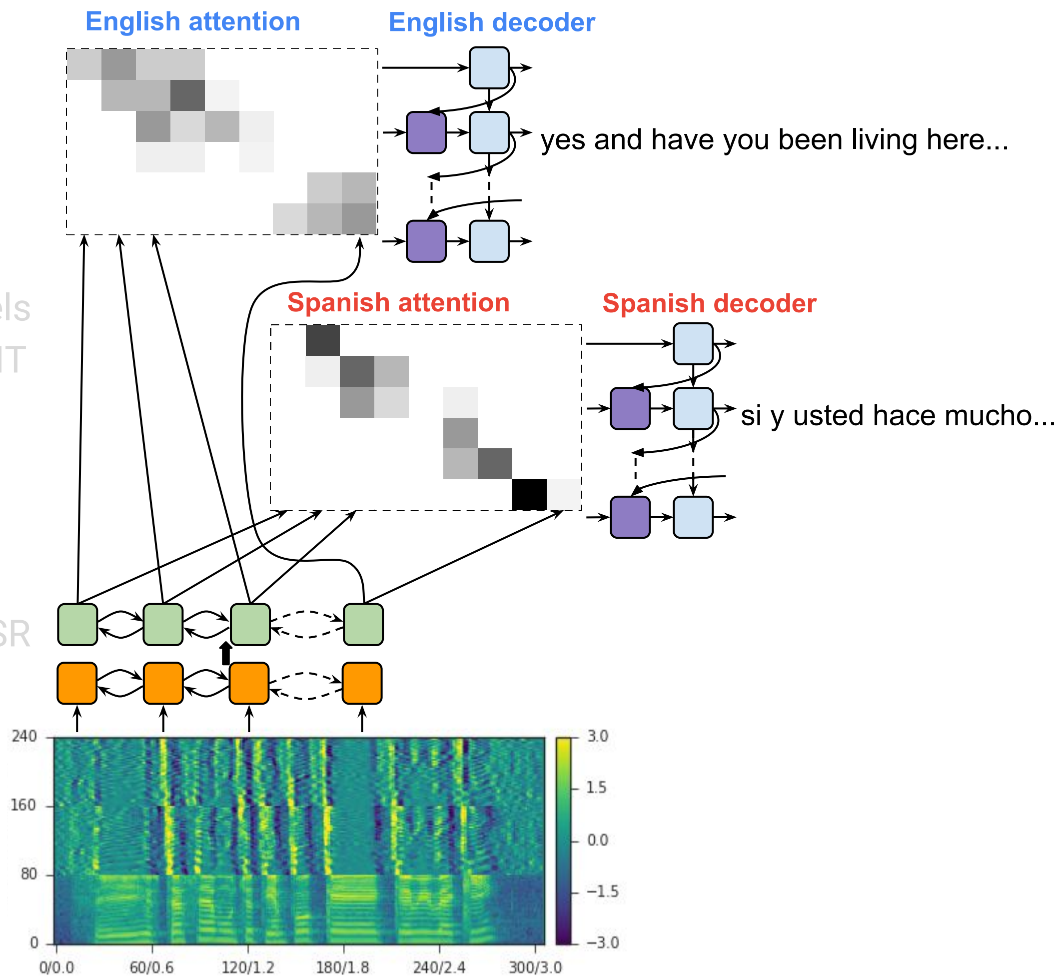
# Seq2seq Speech Translation (ST): Multi-task

Compare three approaches:

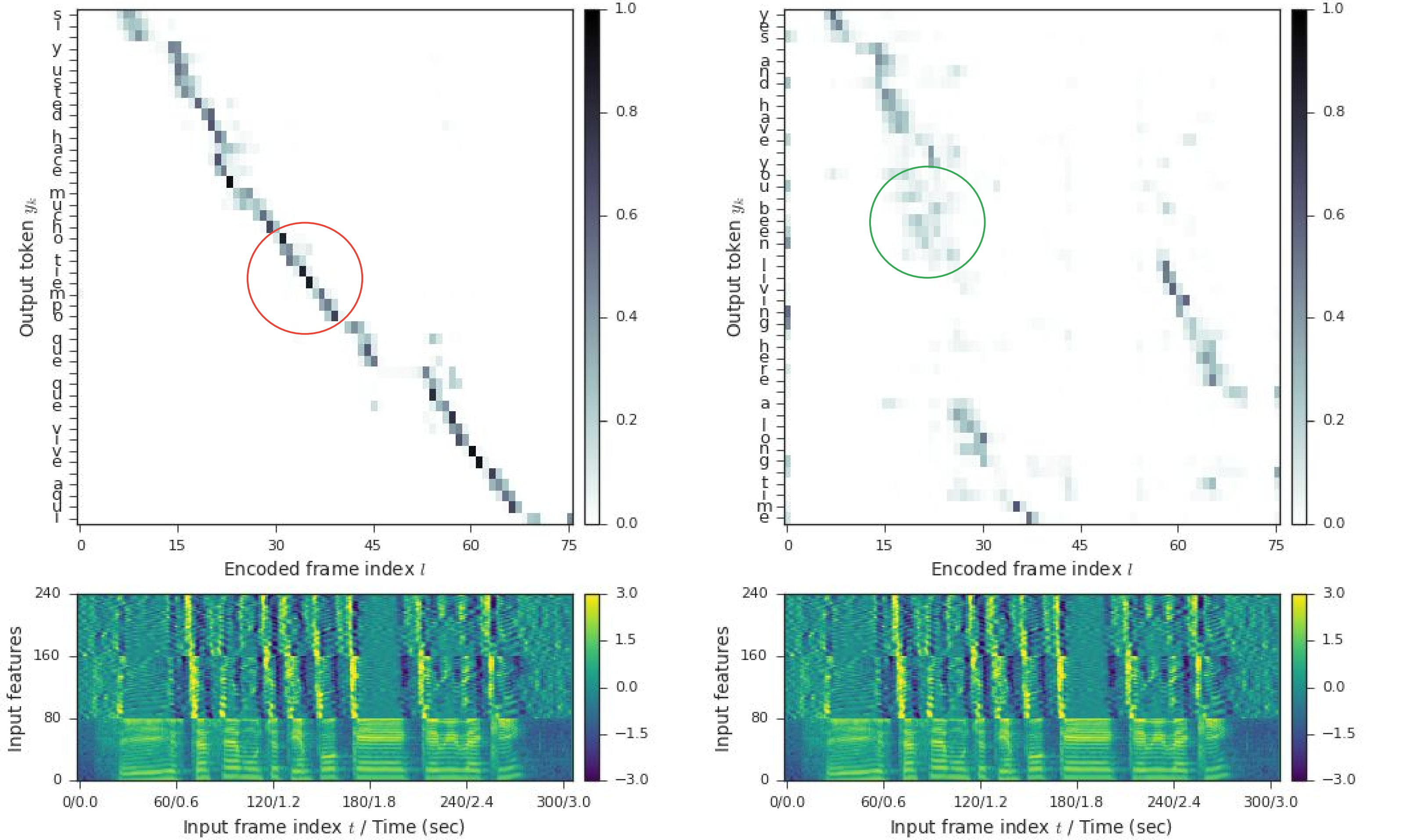
1. ASR -> NMT cascade
   - train independent Spanish ASR, and text neural machine translation models
   - pass top ASR hypothesis through NMT

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



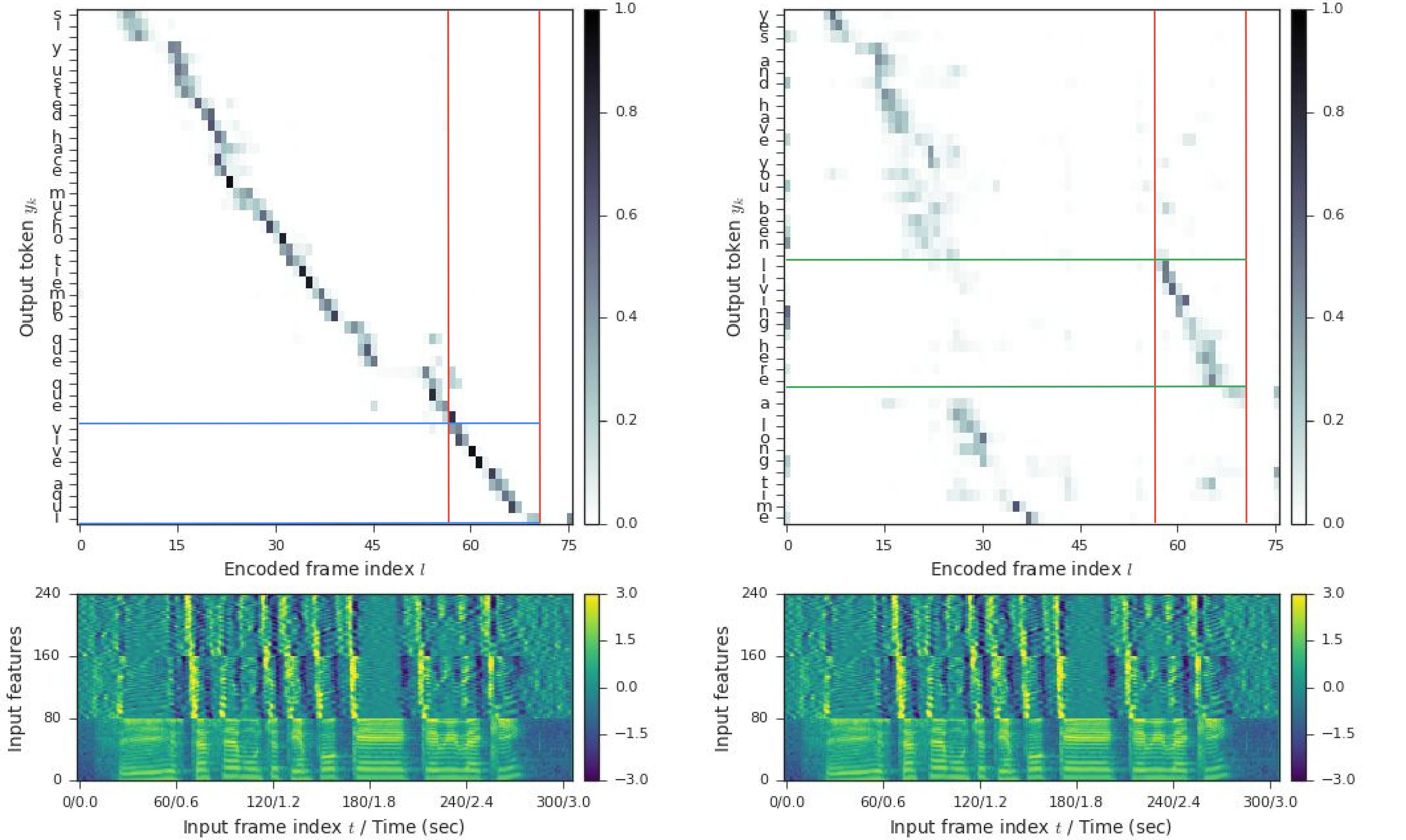
# Seq2seq Speech Translation: Attention




- 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

Google

# 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

| | Fisher | | | Callhome | |
|---|---|---|---|---|---|
| | dev | dev2 | test | devtest | evltest |
| Ours[3] | 25.7 | 25.1 | 23.2 | 44.5 | 45.3 |
| Post et al. [19] | 41.3 | 40.0 | 36.5 | 64.7 | 65.3 |
| Kumar et al. [21] | 29.8 | 29.8 | 25.3 | – | – |

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

| | Fisher | | | Callhome | |
|---|---|---|---|---|---|
| | dev | dev2 | test | devtest | evltest |
| Ours | 58.7 | 59.9 | 57.9 | 28.2 | 27.9 |
| Post et al. [19] | – | – | 58.7 | – | 27.8 |
| Kumar et al. [21] | – | 65.4 | 62.9 | – | – |

- 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

| Model | Fisher dev | Fisher dev2 | Fisher test | Callhome devtest | Callhome evltest |
|---|---|---|---|---|---|
| End-to-end ST[3] | 46.5 | 47.3 | 47.3 | 16.4 | 16.6 |
| Multi-task ST / ASR[3] | 48.3 | 49.1 | 48.7 | 16.8 | 17.4 |
| ASR→NMT cascade[3] | 45.1 | 46.1 | 45.5 | 16.2 | 16.6 |
| Post et al. [19] | – | 35.4 | – | – | 11.7 |
| Kumar et al. [21] | – | 40.1 | 40.4 | – | – |

- 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 really like to dance merengue and salsa"
hyp: "i like to dance merengue and salsa and also"
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


- **[Bérard et al, 2016]** A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *NIPS Workshop on End-to-end Learning for Speech and Audio Processing*, 2016.
- **[Duong et al, 2016]** L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," *NAACL-HLT*, 2016.
- **[Bahdanau et al, 2015]** D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ICLR*, 2015.
- **[Chan et al, 2016]** W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *ICASSP*, 2016.
- **[Chorowski et al, 2015]** J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *NIPS*, 2015.
- **[Zhang et al, 2017]** Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," *ICASSP*, 2017.
- **[Post et al, 2013]** M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus," *IWSLT*, 2013.
- **[Kumar et al, 2015]** G. Kumar, G. W. Blackwood, J. Trmal, D. Povey, and S. Khudanpur, "A coarse-grained model for optimal coupling of ASR and SMT systems for speech translation." *EMNLP*, 2015.

# Extra slides

# End-to-end model: tuning

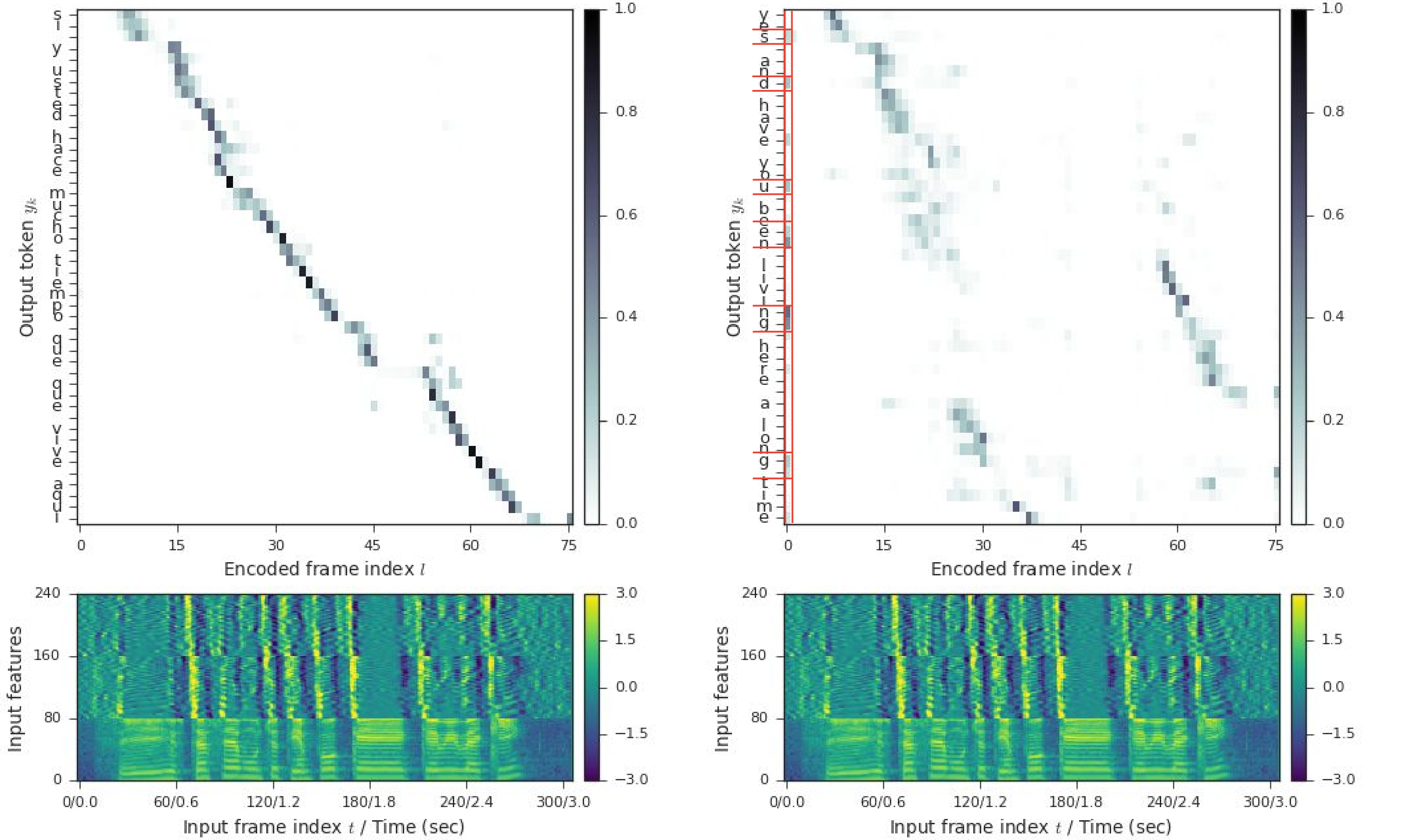
- Performance improves with deeper decoder

| Num decoder layers $D$ | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| 43.8 | 45.1 | 45.2 | 45.5 | 45.3 |

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

| Num shared encoder LSTM layers | | | |
|---|---|---|---|
| 3 (all) | 2 | 1 | 0 |
| 46.2 | 45.1 | 45.3 | 44.2 |

# 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.



Sequence-to-Sequence Models Can Directly Translate Foreign Speech – Interspeech 2017