Google

Generating speech from speech

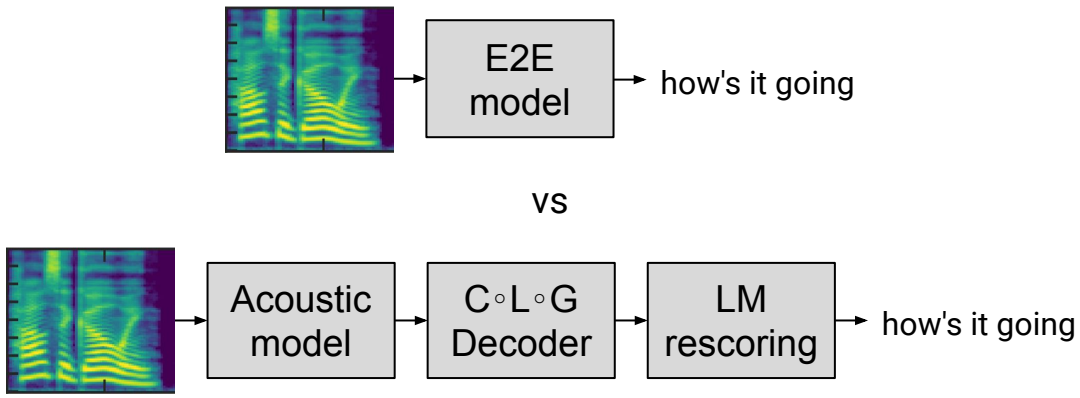
How end-to-end is too far?

Ron Weiss

SANE 2019
2019-10-24

Joint work with Fadi Biadsy, Ye Jia, and many others in Google Brain, Research, Translate, Speech, Project Euphonia

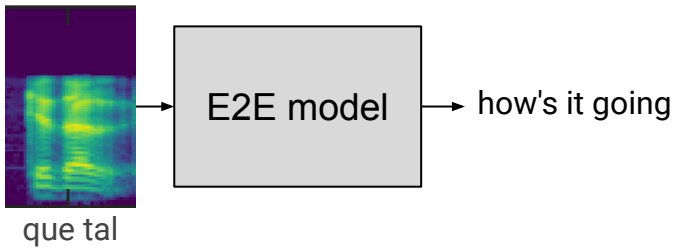
End-to-end / Direct models



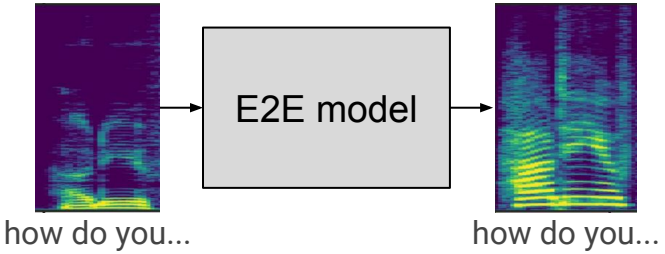
- Single model trained to directly predict desired output from raw input
- Why?
  - Simpler...
    - training data ➔ e.g. don't need lexicon
    - implementation ➔ single neural network, smaller model
    - single decoding step ➔ lower latency inference
  - Directly optimize for desired output
    - avoid compounding errors from intermediate predictions
    - retain useful information in input, e.g. speaker identity, prosody

End-to-end speech models, a journey...

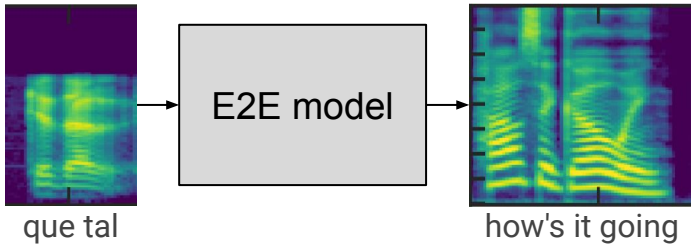
1. 🗣️ **Speech-to-text translation**
   [Weiss, et al., 2017]
   [Jia, Johnson, et al., 2019]



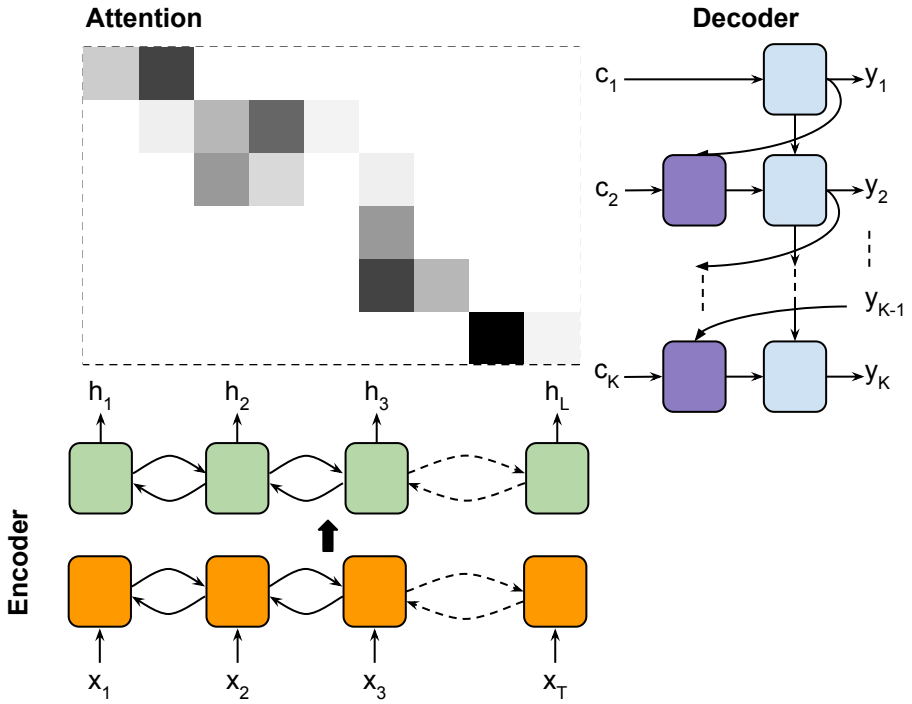
2. 🦜 **Parrotron**
   voice conversion, speech separation
   [Biadsy, et al., 2019]



3. 🐟 **Translatotron**
   speech-to-speech translation
   [Jia, Weiss, et al., 2019]



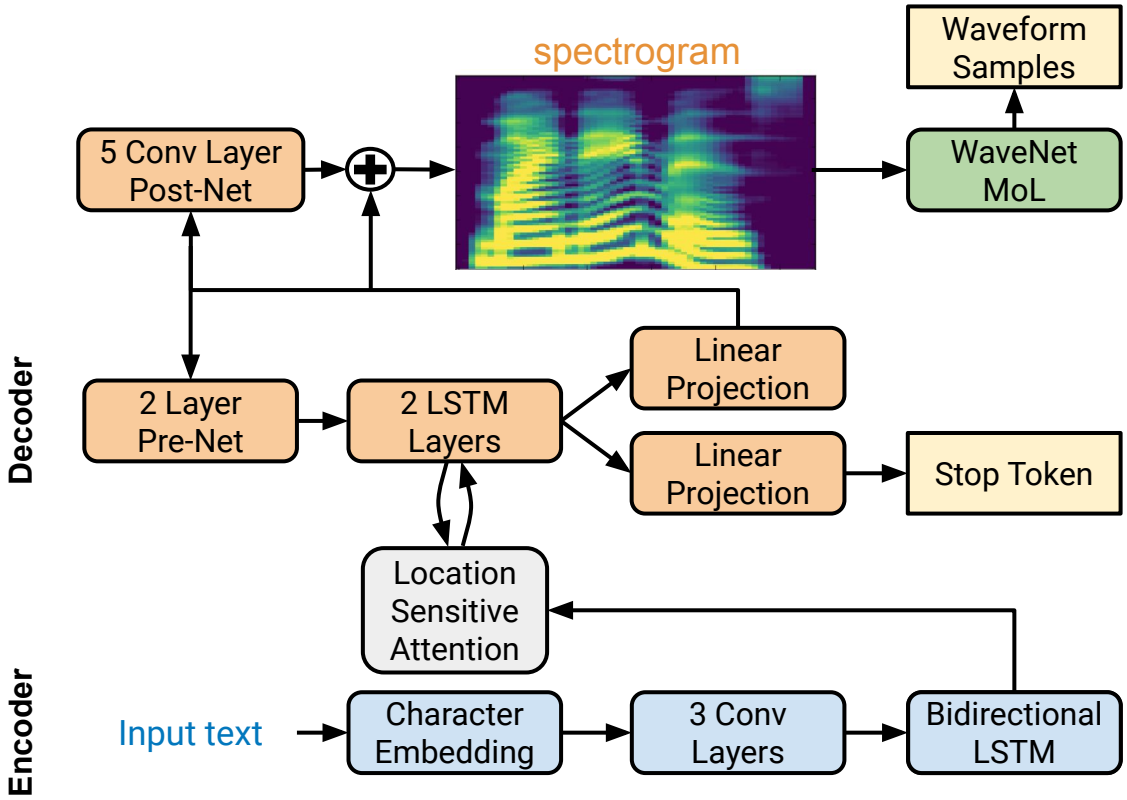
Sequence-to-sequence with attention [Bahdanau et al., 2015]



- Listen, Attend and Spell (LAS) [Chan et al., 2016] [Chorowski et al., 2015]
  - **Encoder** Bidi LSTM stack computes *latent representation* of spectrogram frames
  - **Decoder** Autoregressive LSTM next-step prediction, outputs one character at a time
    - conditioned on entire encoded input sequence via attention context vector
  - **Attention** aligns input and output sequences

Bahdanau, et al., Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015.
Chan, et al., Listen, Attend and Spell. ICASSP 2016.
Chorowski, et al., Attention-Based Models for Speech Recognition. NeurIPS 2015.

Tacotron [Wang et al., 2017] [Shen et al., 2018]



- Inputs are characters, outputs are spectrogram frames
- Separate vocoder network to invert spectrogram to waveform

Wang, et al., Tacotron: Towards End-to-End Speech Synthesis. Interspeech 2017.
Shen, et al., Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. ICASSP 2018.

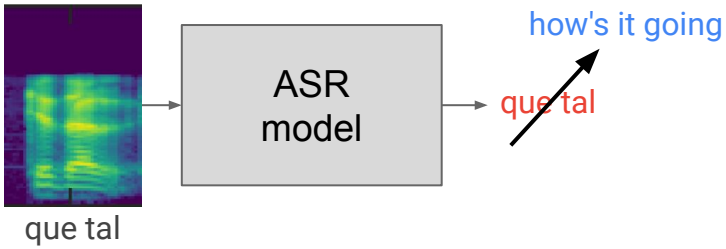
🗣 Speech-to-text translation

Sequence-to-Sequence Models Can Directly Translate Foreign Speech

Ron Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, Zhifeng Chen

Interspeech 2017

Speech-to-text translation (ST)



- Replace the speech transcript with its translation
  - deeper LAS ASR architecture [Zhang et al., 2017]
  - as in "Listen and Translate" on synthetic speech [Bérard et al., 2016]

- Data: Fisher Spanish ➔ English
  - transcribed Spanish telephone conversations from LDC
  - crowdsourced English translations of Spanish transcripts from [Post et al., 2013]
  - train on 140k Fisher utterances (160 hours)

Zhang, et al., Very Deep Convolutional Networks for End-to-End Speech Recognition. ICASSP 2017
Bérard, et al., Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. NeurIPS 2018.
Post, et al., Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus. IWSLT 2013

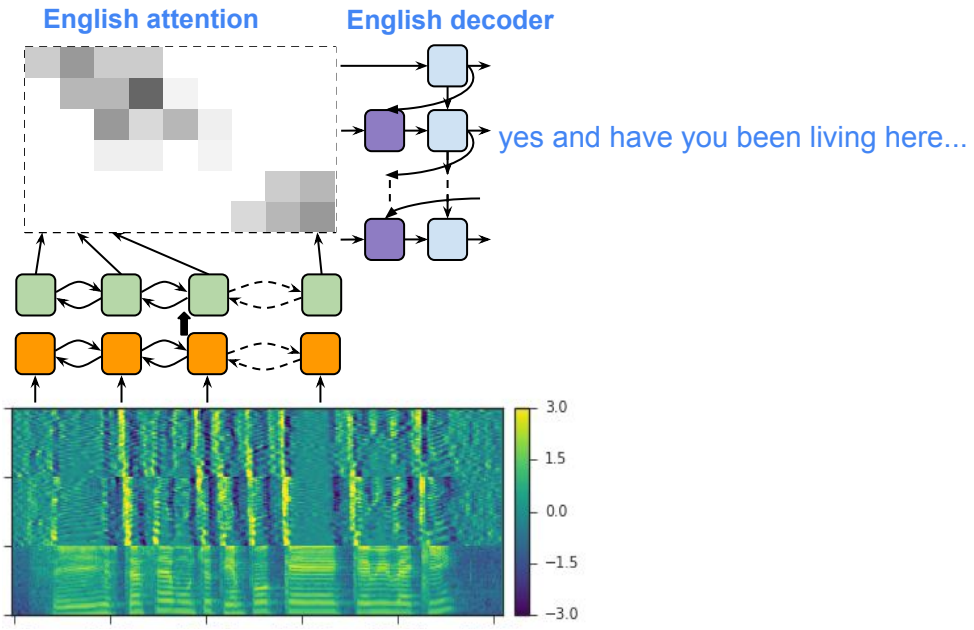
ST models: End-to-end

Compare three approaches:

1. **End-to-end ST**
   - train LAS model to directly predict *English* text from Spanish audio

2. Multi-task ST / ASR
   - *shared* encoder
   - 2 independent decoders with *different* attention networks, each emitting text in a different language

3. ASR ➔ NMT cascade
   - train independent Spanish ASR, and text neural machine translation models
   - pass top ASR hyp through NMT



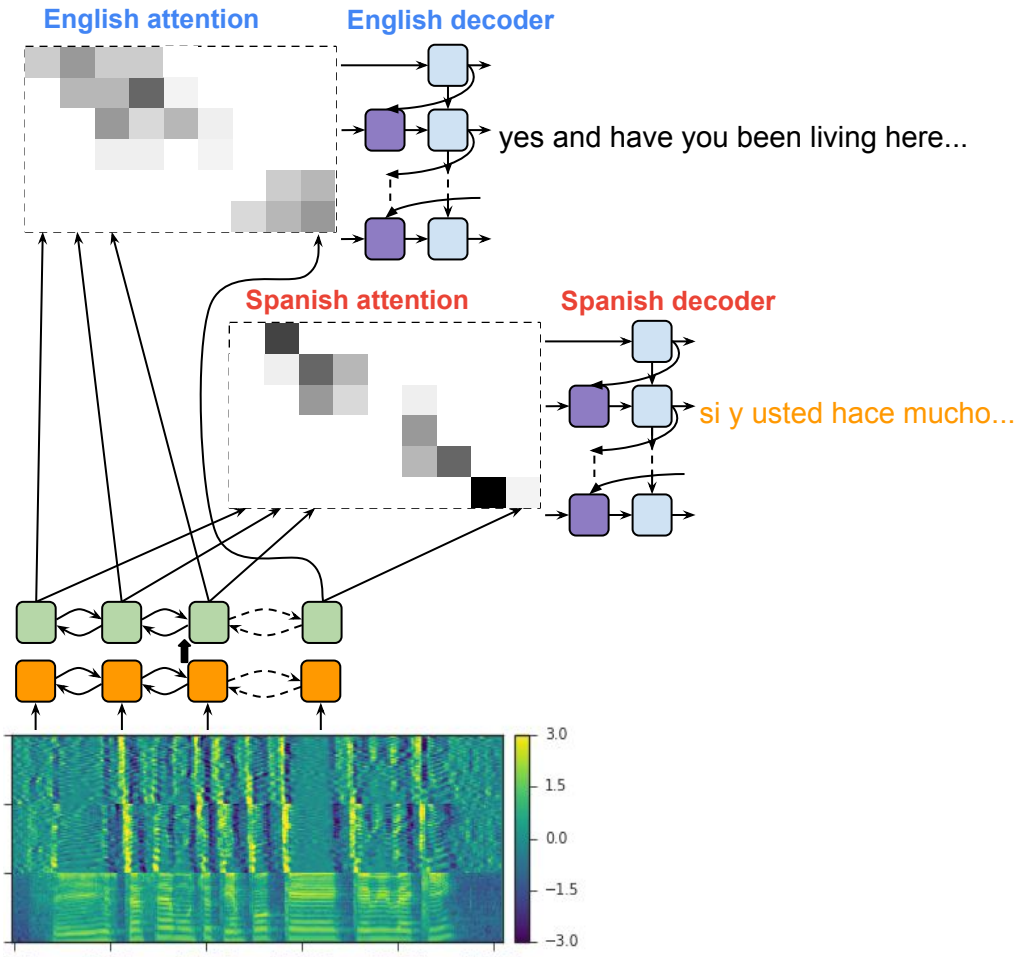
ST models: Multi-task

Compare three approaches:

1. End-to-end ST
   - train LAS model to directly predict *English* text from Spanish audio

2. **Multi-task ST / ASR**
   - *shared* encoder
   - 2 independent decoders with different attention networks, each emitting text in a different language

3. ASR ➔ NMT cascade
   - train independent Spanish ASR, and text neural machine translation models
   - pass top ASR hyp through NMT



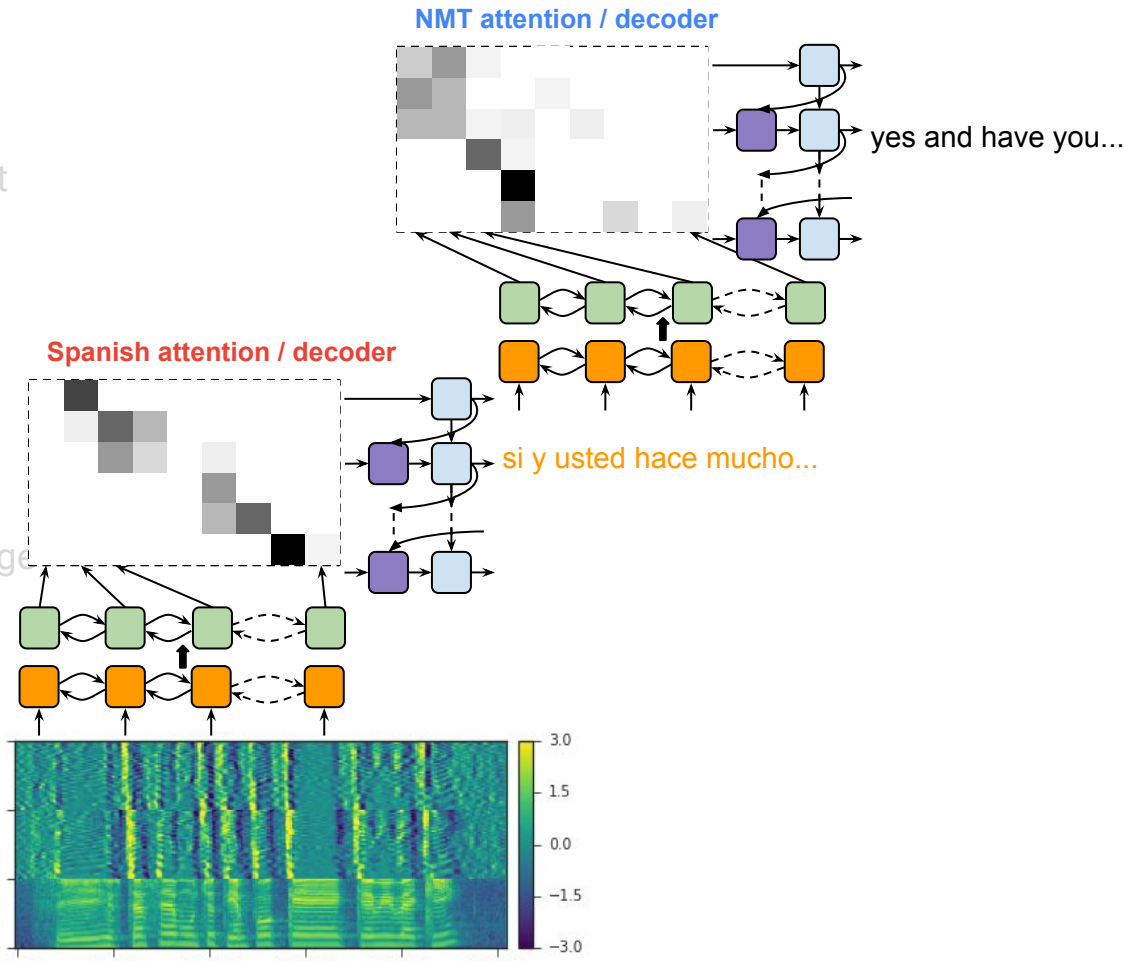
ST models: Cascade

Compare three approaches:

1. End-to-end ST
   - train LAS model to directly predict *English* text from Spanish audio

2. Multi-task ST / ASR
   - *shared* encoder
   - 2 independent decoders with *different* attention networks, each emitting text in a different language

3. **ASR ➔ NMT cascade**
   - train independent Spanish ASR, text neural machine translation
   - pass top ASR hyp through NMT



Results

Model	dev	Fisher dev2	Fisher test
End-to-end ST[3]	46.5	47.3	47.3
Multi-task ST / ASR[3]	48.3	49.1	48.7
ASR→NMT cascade[3]	45.1	46.1	45.5
Post et al. [19]	–	35.4	–
Kumar et al. [21]	–	40.1	40.4

- BLEU score (higher is better)
- **Multi-task** > End-to-end ST > Cascade >> non-seq2seq baselines

🗣 Speech-to-text translation

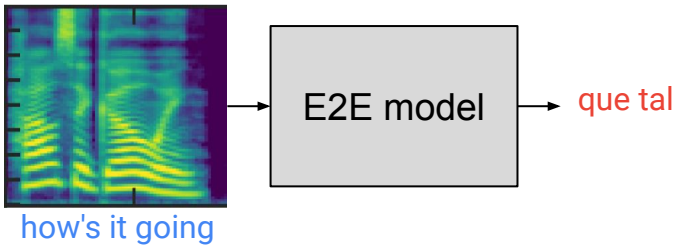
Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, Yonghui Wu

ICASSP 2019

Scaling up

- ST1: 1M utterances, *read* English speech ➔ Spanish text
  - conversational speech translation
  - 1k+ hours, ~7x larger than Fisher

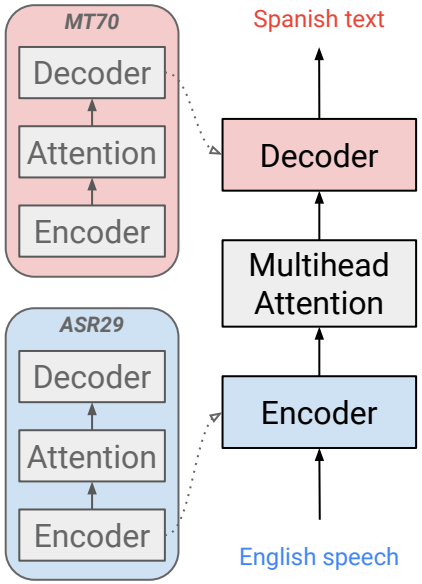


- Compare with baseline English ASR ➔ English-to-Spanish NMT cascade
  - **ASR29**: 29M transcribed English utterances
    - anonymized voice search logs, 40k+ hours
  - **MT70:** 70M English text ➔ Spanish text
    - web text, superset of ST1

Scaled up model

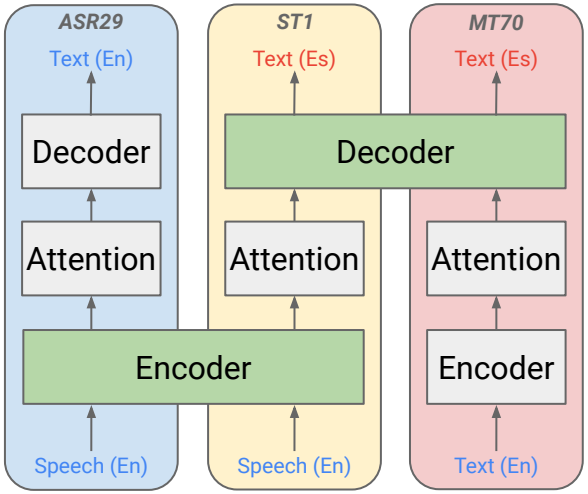
- Bigger model
  - 5x BLSTM encoder, 8x LSTM decoder, 8-head attention
  - 62M parameters, ~6x larger than Fisher ST
- Underperforms cascade of ASR and NMT models
  - but trained on 100x fewer examples
- Pretraining bridges most of the gap
  - encoder pretrained on ASR29
  - decoder pretrained on MT70

Model	Fine-tuning set	BLEU
Cascaded		**56.9**
Fused	ST1	49.1
Fused + pretrained	ST1	54.6



Weak supervision: Multi-task training

- Multi-task train with all available data
  - share encoder with English ASR
  - share decoder with English ➔ Spanish NMT
  - sample task independently at each step



- Matches cascade performance

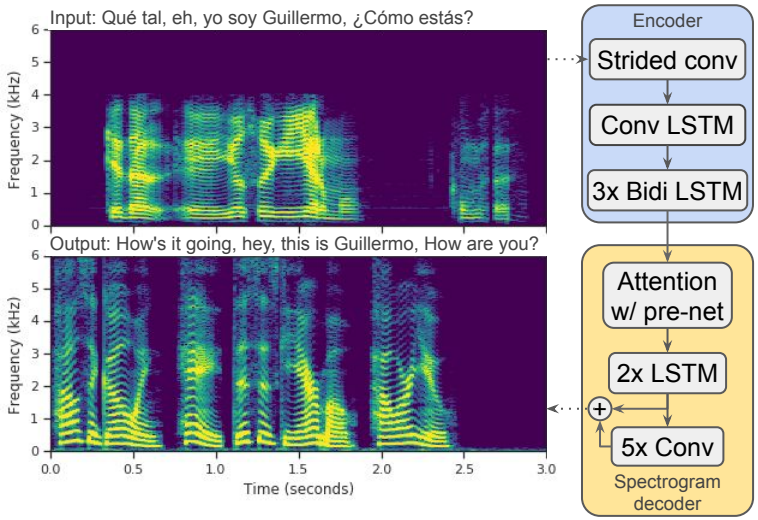
Model	Fine-tuning set	BLEU
Cascaded		56.9
Fused	ST1	49.1
Fused + pretrained	ST1	54.6
Fused + pretrained + multi-task	ST1	**57.1**

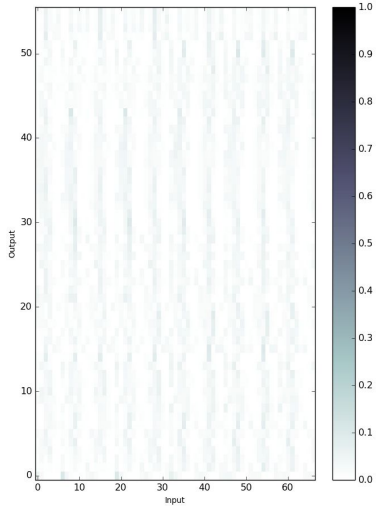
Interlude:
Speech-to-speech Translation, Take 1

Toward direct speech-to-speech translation

- Join ST encoder with Tacotron 2 decoder
- Fisher Spanish ➔ English (TTS)
  - synthesize target English speech from translated transcript using single speaker Parallel WaveNet TTS [van den Oord et al., 2018]
- Fails to learn attention, just babbles

Output Target

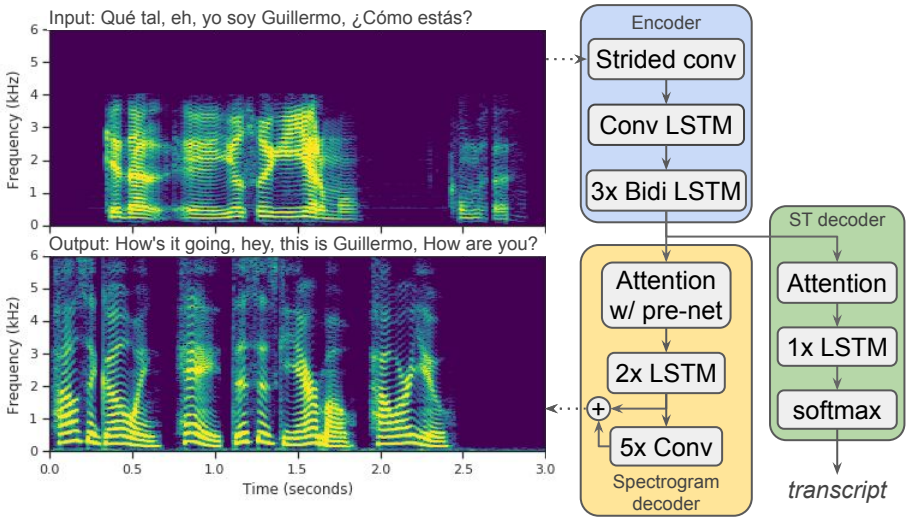




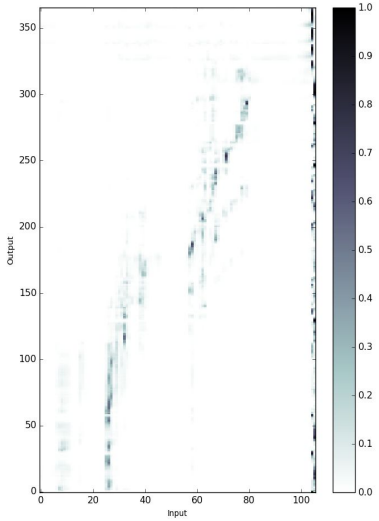
van den Oord, et al., Parallel WaveNet: Fast High-Fidelity Speech Synthesis. ICML 2018.

Multi-task training

- Add auxiliary decoder
  - predict output (translated) transcript
  - helps encoder learn more useful internal representation
    - not used during inference
  - ~35M parameters



- Begins to pick up attention
  - but only learns to translate common words, short phrases



Output Target

Output Target

Voice conversion

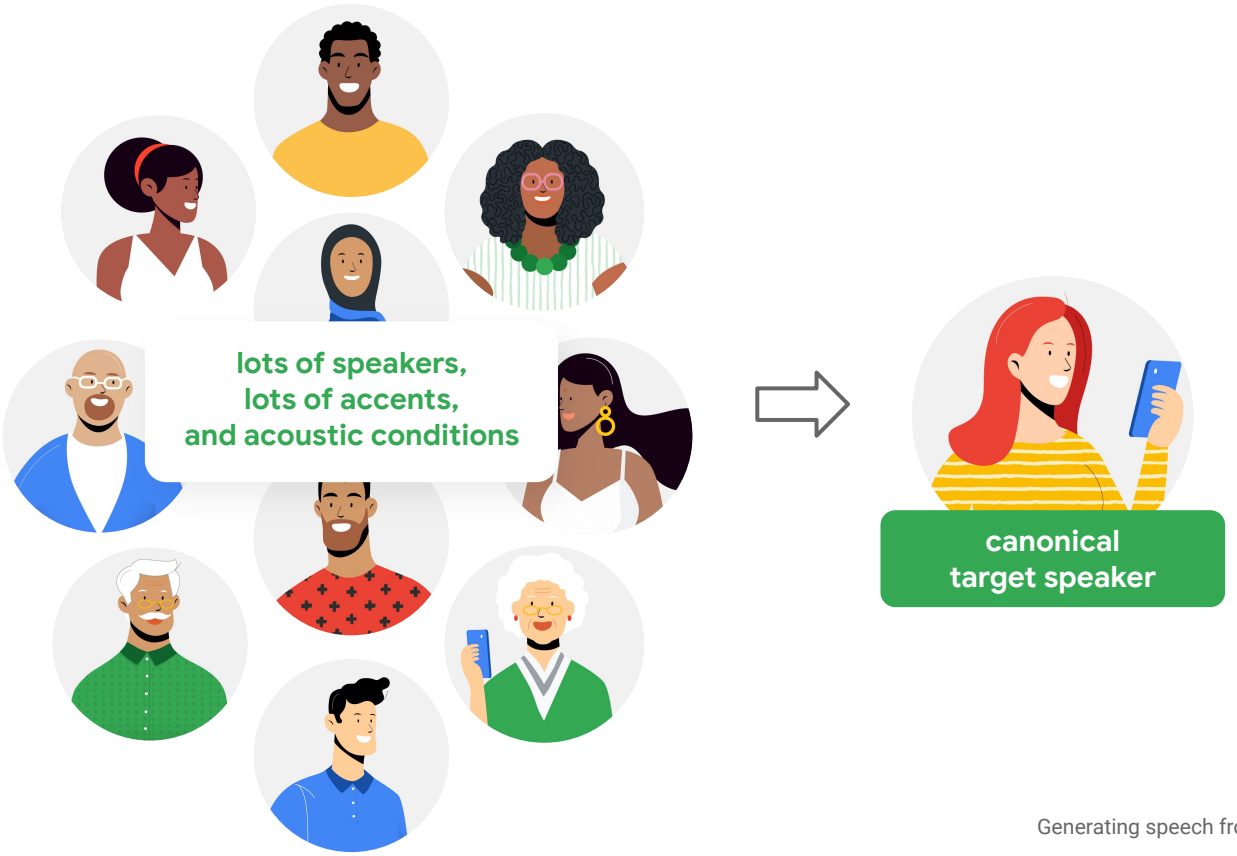
Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation

Fadi Biadsy, Ron Weiss, Pedro Moreno, Dimitri Kanevsky, Ye Jia

Interspeech 2019

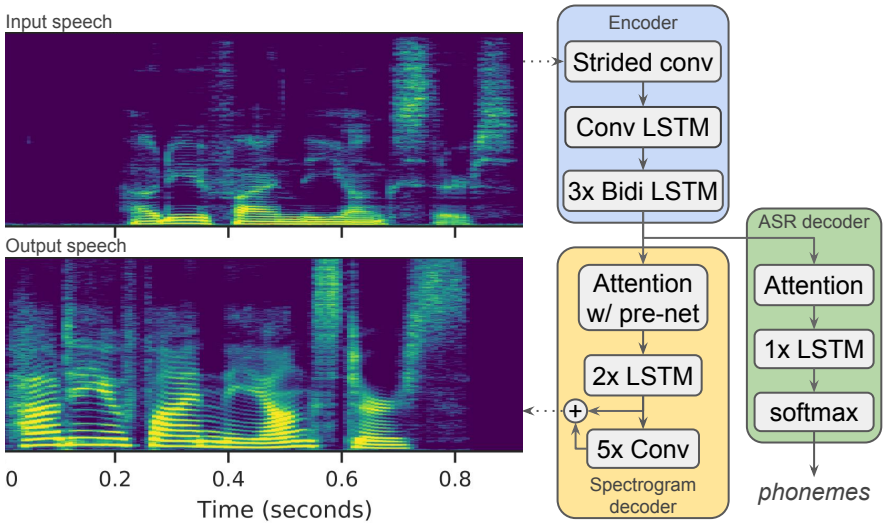
Voice normalization

- Voice normalization = text-independent, many-to-one voice conversion to *canonical* target voice
- Retain what is being said, not who, where, or how
  - i.e. discard speaker identity, accent, prosody, background noise, ...

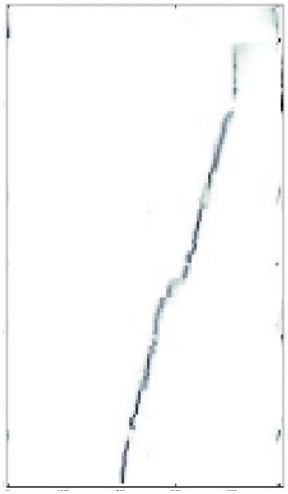


Parrotron

- Same architecture:
  ASR encoder ➔ Tacotron 2 decoder
  multi-task with ASR decoder



- Train on ASR29
  - real speech input
  - synthesize target speech from transcript using single speaker Parallel WaveNet TTS [van den Oord et al., 2018]



van den Oord, et al., Parallel WaveNet: Fast High-Fidelity Speech Synthesis. ICML 2018.

Voice normalization results

ASR decoder target	#CLSTM	#LSTM	Attention	WER
None	1	3	Additive	27.1
Grapheme	1	3	Location	19.2
Phoneme	1	3	Location	18.5
Phoneme	0	5	Location	18.3
Phoneme w/slow decay	0	5	Location	17.6

- Evaluate with ASR word error rate (WER)
- Multi-task training important
  - phoneme targets perform better than graphemes

Input Output

Input Output

Input Output

Source: https://google.github.io/tacotron/publications/parrotron

Atypical speech conversion

- **Task:** Convert “atypical” speech to speech that can be more easily understood by both humans and off-the-shelf speech systems
  - e.g. due to deafness or physical/neurological conditions such as ALS, multiple sclerosis, ...

- Case study: speech from a deaf speaker
  - profoundly deaf since the age of 1
  - born in Russia, learned English as a teenager

Source: https://www.youtube.com/watch?v=Act4NIe-sBg

I am Dimitri Kanevsky.
I'm a speech research scientist at Google.
Before Google I worked at IBM.

Atypical speech conversion: Results

- Adapt normalization model using 13 hours of transcribed speech

Input Output After adaptation

Input Output After adaptation

Source: https://google.github.io/tacotron/publications/parrotron

Model	MOS	WER
Real speech	2.08 ± 0.22	89.2
Parrotron (male)	2.58 ± 0.20	109.3
Parrotron (male) finetuned	$\mathbf{3.52 \pm 0.14}$	**32.7**

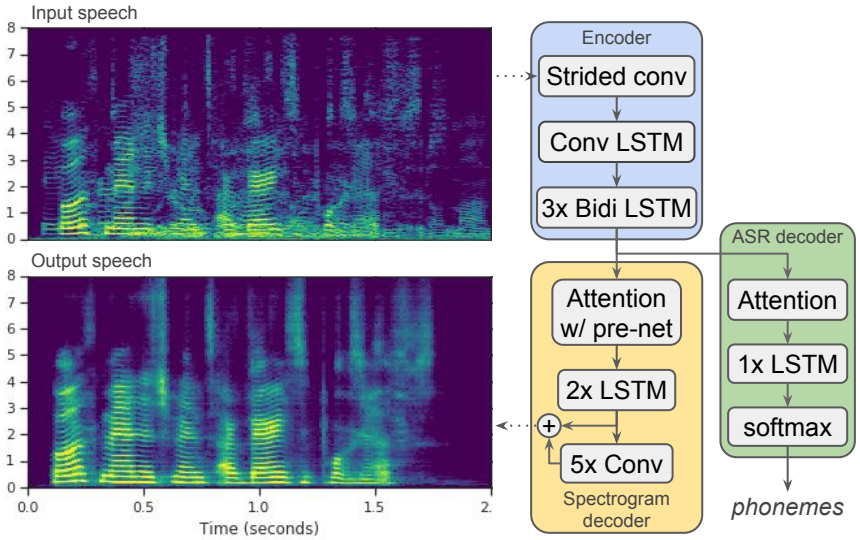
- Substantially improves subjective naturalness and intelligibility

Atypical speech conversion: Demo

Source: https://www.youtube.com/watch?v=KtKGWSpppz4

Speech separation by synthesis

- **Task** Extract loudest speaker in mixture
  - up to 8 overlapping speakers
  - instantaneous mixing
  - 12 dB avg SNR



Data	WER
Original (Clean)	8.8
Noisy	33.2
Denoised using Parrotron	**17.3**

- Large improvement in intelligibility
- Same architecture can be used to generate real speech from many speakers

Input Output

Input Output

Source: https://google.github.io/tacotron/publications/parrotron

🐟 Speech-to-speech Translation, Take 2

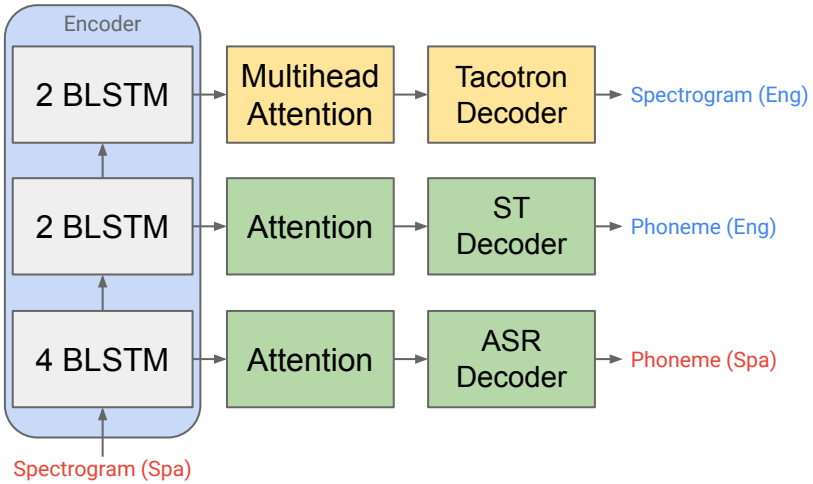
Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model

Ye Jia, Ron Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, Yonghui Wu

Interspeech 2019

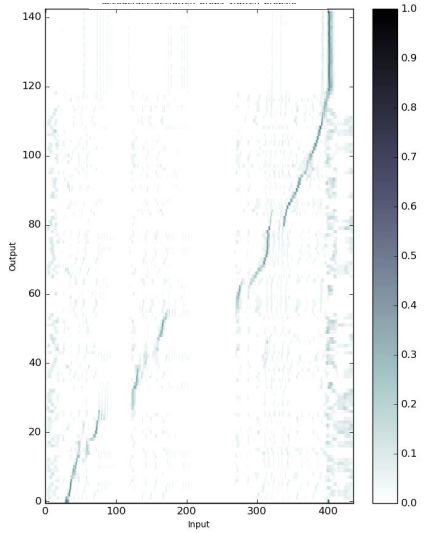
Translatotron: Making speech-to-speech translation work

- Apply lessons from ST and Parrotron
  - deeper model
    - 8x BLSTM encoder, 4x LSTM decoder
    - 69M parameters
  - 4-head attention
  - multi-task training
    - predict source and target phonemes from intermediate encoder outputs
  - encoder pretraining

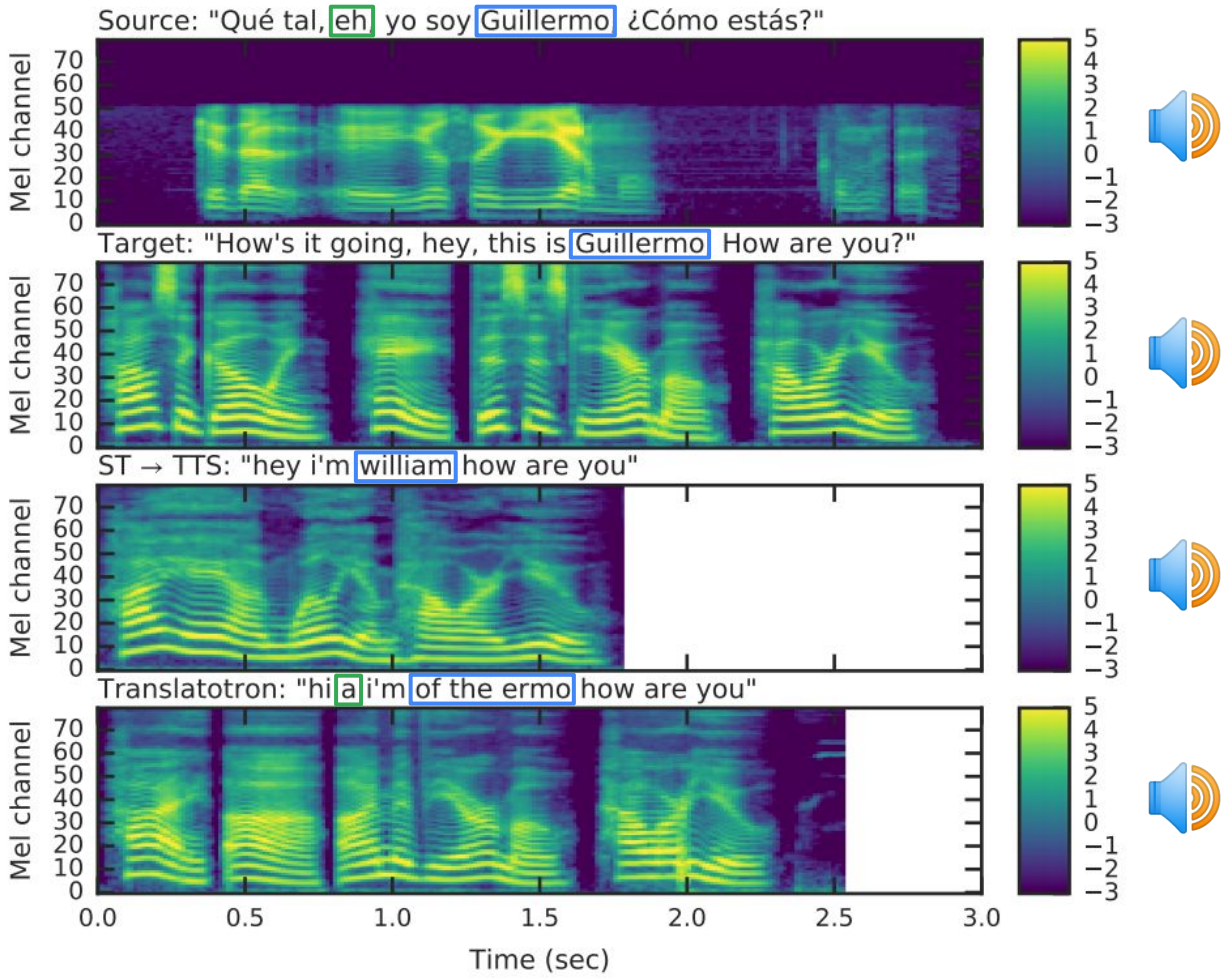


- Train on Fisher Spanish ➔ English (TTS)

- Other tricks to pick up attention
  - narrow pre-net bottleneck in autoregressive path
  - fewer output steps:
    predict two frames per decoder step

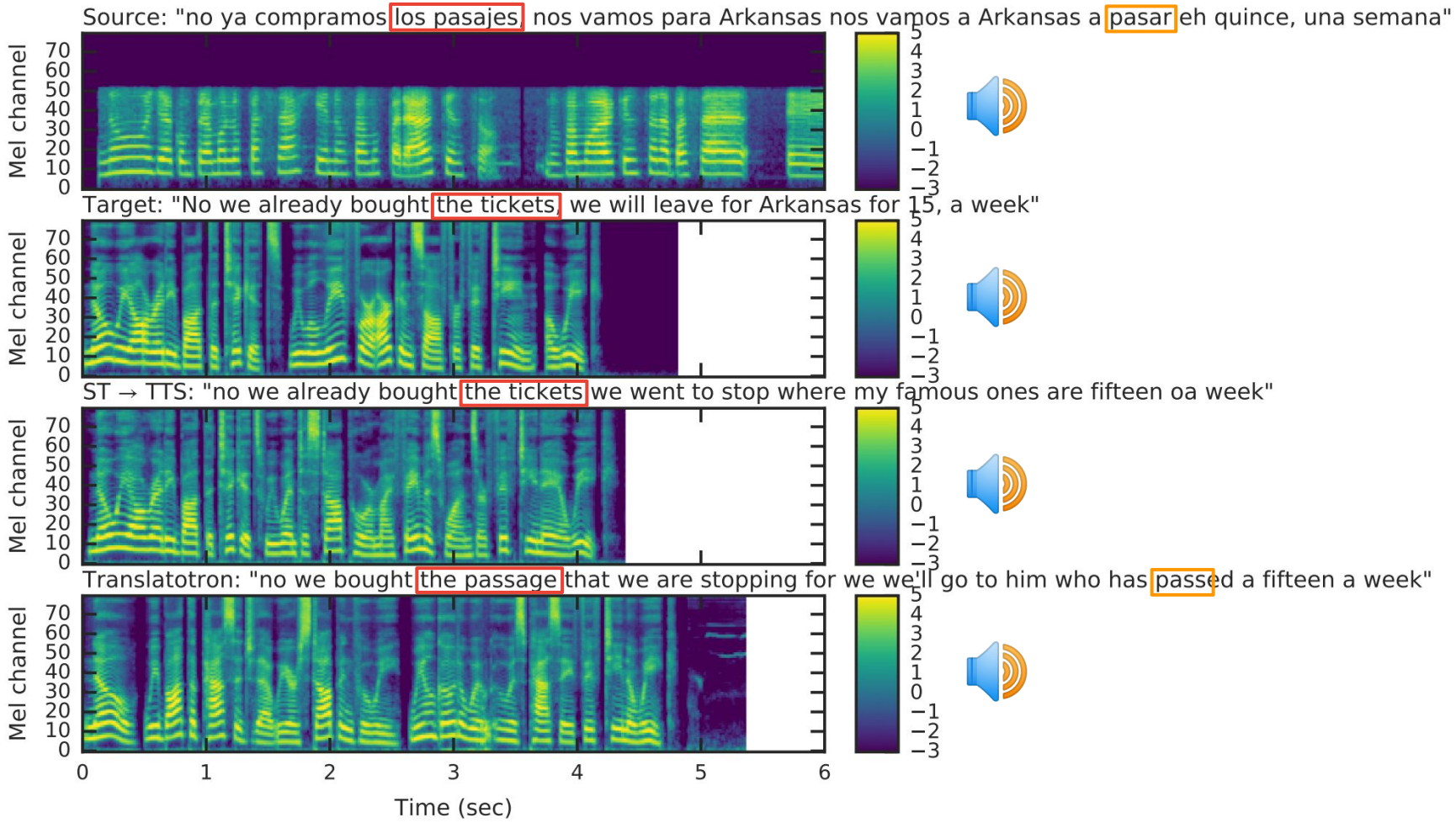


Samples



Source: https://google-research.github.io/lingvo-lab/translatotron/#fisher_1

Samples 2



Source: https://google-research.github.io/lingvo-lab/translatotron/#fisher_7

Performance

- Measuring translation quality
  - transcribe output speech using LibriSpeech recognizer [Irie et al., 2019]
  - compute BLEU on output ➔ lower bound due to ASR errors

Auxiliary loss	dev1	dev2	test
None	0.4	0.6	0.6
Source	7.4	8.0	7.2
Target	20.2	21.4	20.8
Source + Target	24.8	26.5	25.6
Source + Target (1-head attention)	23.0	24.2	23.4
Source + Target (encoder pre-training)	30.1	31.5	31.1
ST [19] → TTS cascade	39.4	41.2	41.4
Ground truth	82.8	83.8	85.3

- Multi-task training to predict target phonemes is critical
- But helps to predict target and source *and* pretrain the encoder
- Underperforms cascade baseline
- Same trend in subjective listening tests

Model	Fisher-test
Translatotron	3.69 ± 0.07
ST→TTS	4.09 ± 0.06

Irie, et al., On the Choice of Modeling Unit for Sequence-to-Sequence Speech Recognition. Interspeech 2019.

Translatotron on ST1

- Train on ST1 Spanish ➔ English
  - synthesize target English speech from translation using Tacotron 2 [Shen et al., 2018]
- Wider layers, deeper (6 layer) decoder
  - ~5x larger model than Fisher, 320M parameters

Auxiliary loss	BLEU	Source PER	Target PER
None	0.4	-	-
Source	42.2	5.0	-
Target	42.6	-	20.9
Source + Target	42.7	5.1	20.8
ST [21] → TTS cascade	48.7	-	-
Ground truth	74.7	-	-

- All you need is... multi-task training
- Still underperforms cascade

Model	Conversational
Translatotron	4.08 ± 0.06
ST→TTS	4.32 ± 0.05

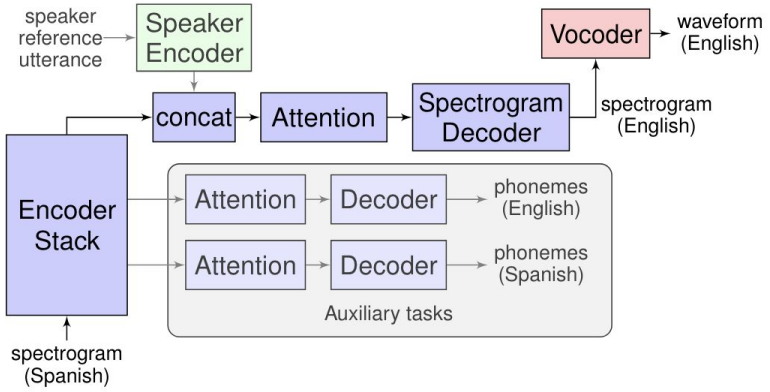
Shen, et al., Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. ICASSP 2018.

Cross-lingual voice transfer

- ST1 subset with Spanish *and* English speech
  - ~600k utterances
  - but mismatched speakers

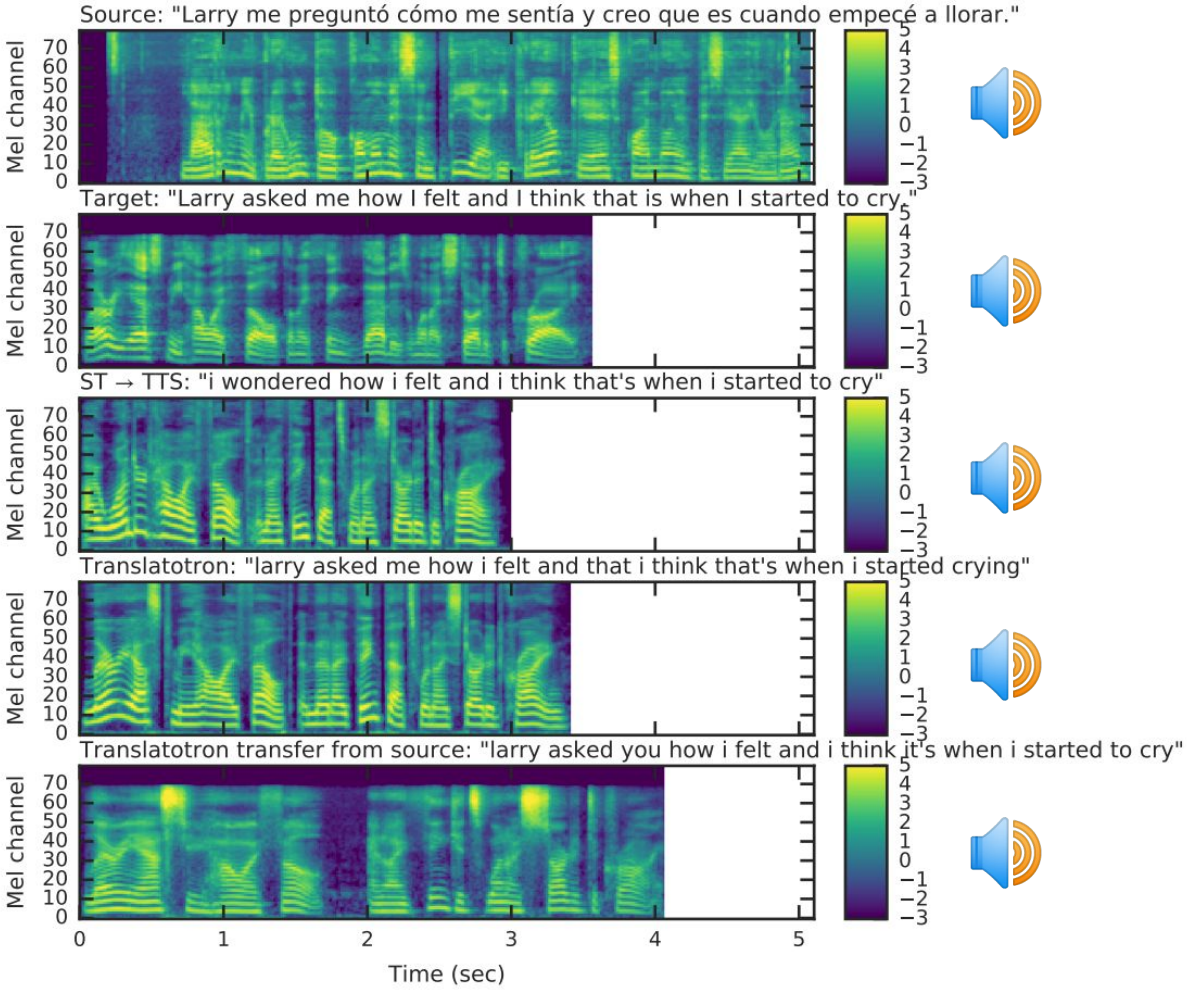
- Condition decoder on **speaker embedding** computed from reference utterance
  - use target English utterance during training
  - during inference, use source Spanish utterance
  - similar to zero-shot speaker adaptation for TTS [Jia et al., 2018]

- Pretrain (and freeze) speaker encoder on speaker verification task [Wan et al., 2018]
  - trained on 8 languages including English and Spanish



Jia, et al., Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. NeurIPS 2018.
Wan, et al., Generalized End-to-End Loss for Speaker Verification. ICASSP 2018.

Voice transfer: Samples



Source: https://google-research.github.io/lingvo-lab/translatotron/#conversational_9

Voice transfer: Performance

Speaker Emb	BLEU	MOS-naturalness	MOS-similarity
Source	33.6	3.07 ± 0.08	1.85 ± 0.06
Target	36.2	3.15 ± 0.08	3.30 ± 0.09
Random target	35.4	3.08 ± 0.08	3.24 ± 0.08
Ground truth	59.9	4.10 ± 0.06	-

- Worse than non-voice transfer model
  - due to noisier training targets, smaller training set
- Condition on reference utterances from different speakers
  - source utterance (cross-language voice transfer) worse than English speaker
  - speaker encoder is sensitive to language
    - mismatched in training

Summary

Sequence-to-sequence models are pretty powerful...

- 🗣️ Speech-to-text translation
  - repurpose ASR architecture
  - incorporate weak supervision from other data
    - multi-task training improves performance
- 🦜 Voice conversion
  - ASR encoder + TTS decoder
  - train to predict synthetic speech targets
  - adapt to atypical speech
- 🐟 Speech-to-speech translation
  - requires multi-task training
    - but no text representation used during inference
  - underperforms cascade, but better at maintaining prosody
  - transfer source voice to translated speech

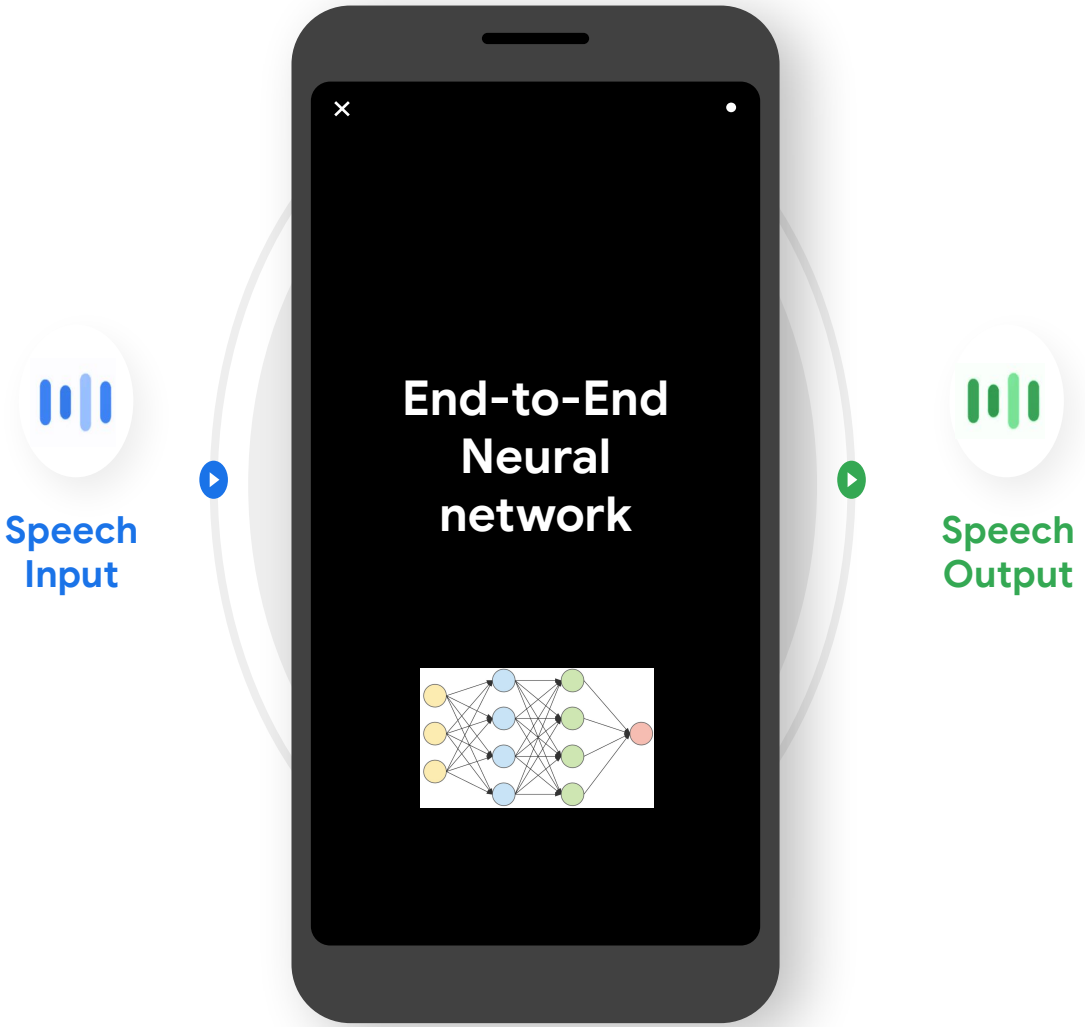
End-to-end food for thought / Have we gone too far?

- Why not end-to-end?
  - More difficult to collect training data at scale
  - Different biases
    - e.g. Parrotron copies pronunciation? Translatotron prefers cognates?
  - Less interpretable
    - cascade's intermediate hyps are useful

- Are direct models less powerful?
  - Limited to a single decode / beam search
    - how to represent multimodality or uncertainty?
  - Tacotron decoder does greedy decoding
    - why multi-task training is critical for Translatotron?

Thanks!

Live demo

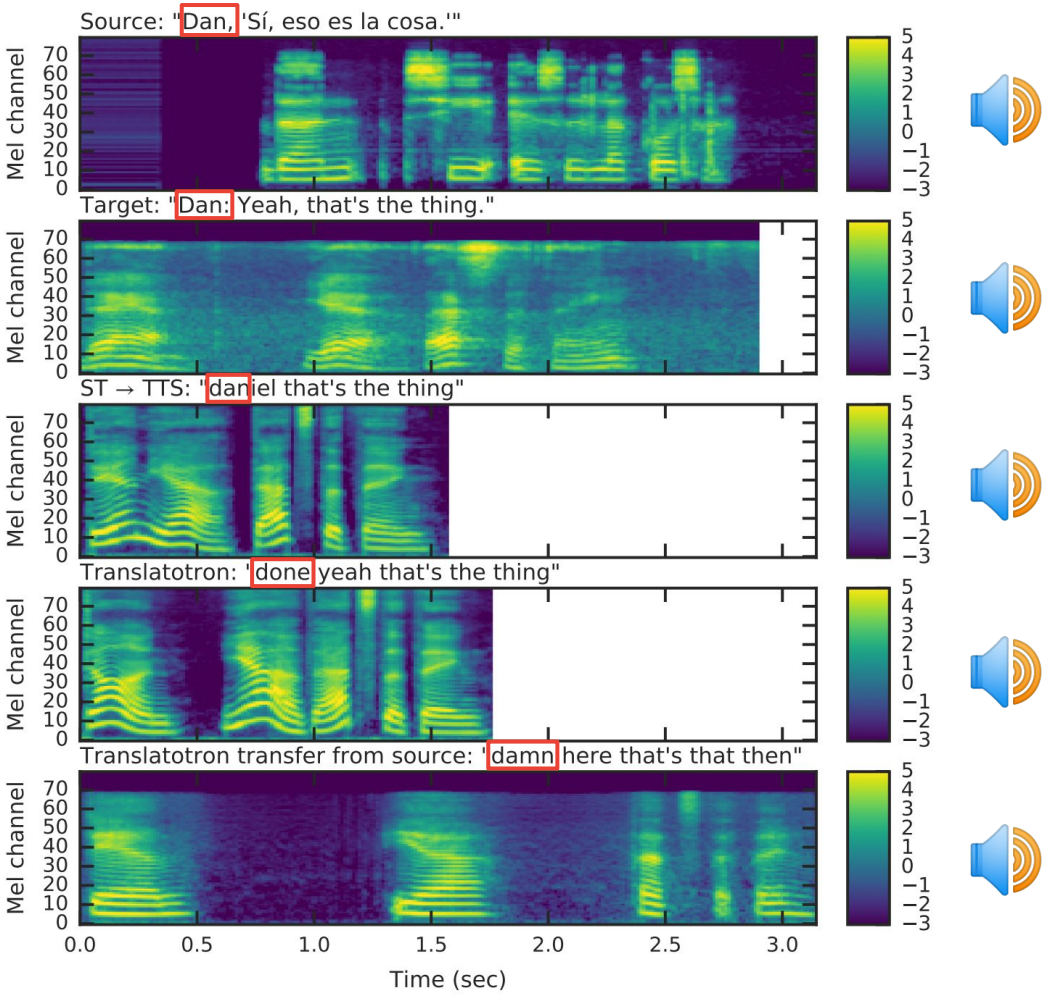


Extra slides

Atypical speech conversion demo 2

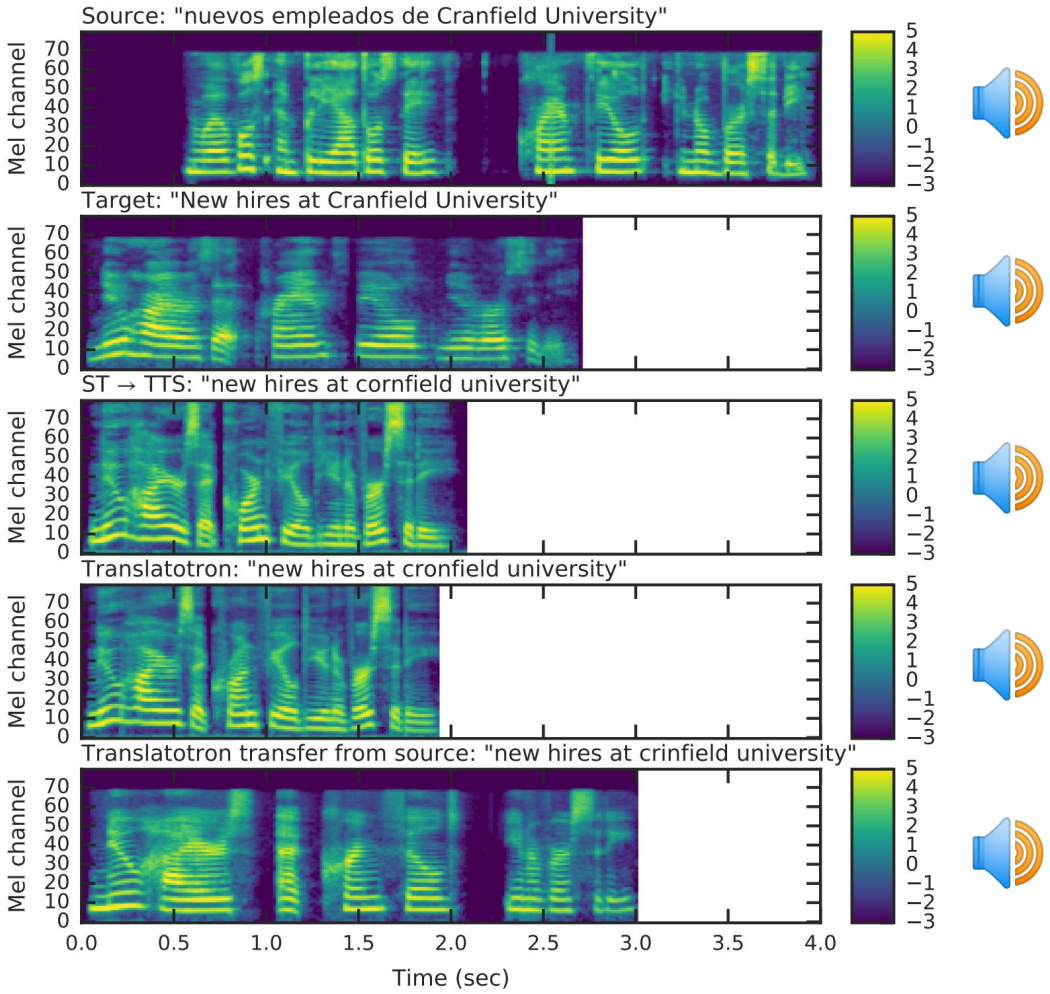
Source: https://www.youtube.com/watch?v=Act4NIe-sBg

Voice transfer: Samples 2



Source: https://google-research.github.io/lingvo-lab/translatotron/#conversational_2

Voice transfer: Samples 3



Source: https://google-research.github.io/lingvo-lab/translatotron/#conversational_5