Google

Translatotron 2: High-quality direct speech-to-speech translation with voice preservation

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, Roi Pomerantz

Google Research

ICML 2022

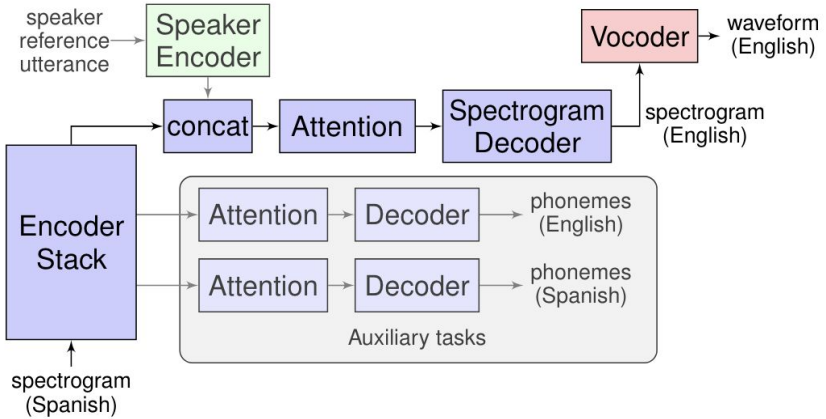
Background – Speech-to-speech translation (S2ST)

- Conventional S2ST (cascade systems)

- Direct S2ST (Translatotron, etc.)

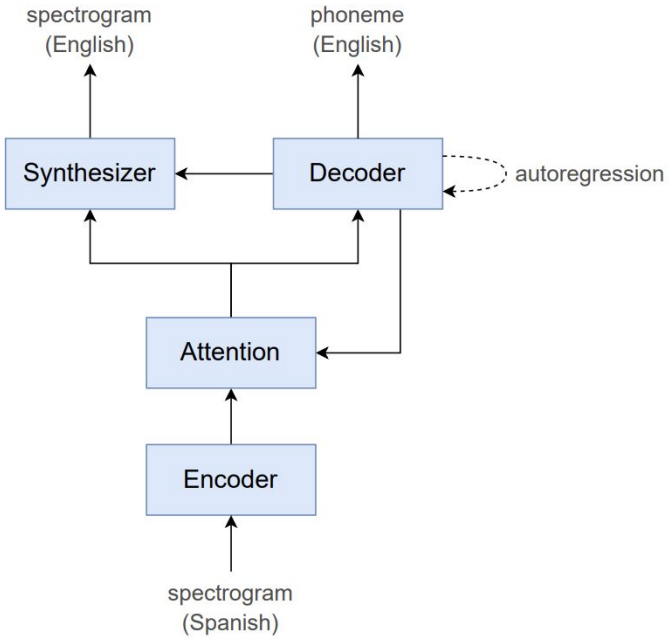
Background – Translatotron (2019)

- First direct S2ST model
  - seq2seq w/ attention
  - Multi-task trained
  - Voice transfer (trained on real world data)
- Limitation
  - Significant translation quality gap to conventional cascade systems
  - Robustness issues (e.g. babbling, early-stopping)



Translatotron 2 – Overview

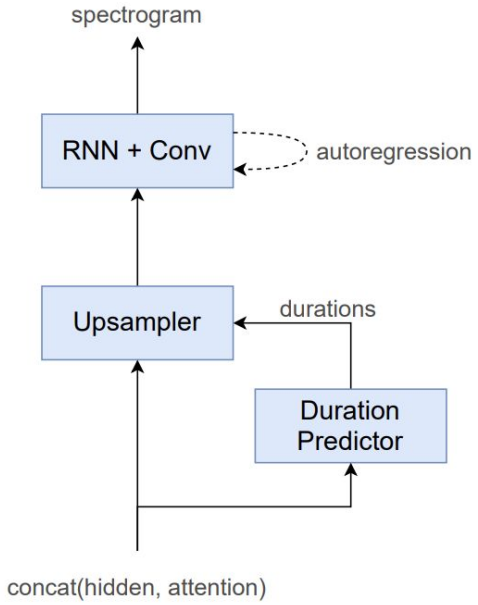
- Splits the original decoder to:
  - A linguistic decoder
  - An acoustic synthesizer
- A single attention connects three components
  - Inputs to the decoder and the synthesize is temporally synchronized
- Attention is driven by the linguistic decoder
  - Avoids directly learning alignment between long spectrogram sequences
  - Textual supervision is better utilized during training



(a) Overview of Translatotron 2.

Translatotron 2 – Synthesizer

- Non-attentive architecture
  - Avoids robustness issues in attention-based speech generation
- Duration predictor is learned without direct supervision



(b) Synthesizer of Translatotron 2.

Performance

Performance on Fisher Spanish-English

	BLEU	Naturalness (MOS)
Translatotron 2	**42.4**	**3.98 + 0.08**
Translatotron	26.9	3.70 + 0.08
Cascade (ST + TTS)	43.3	4.04 + 0.08
Reference (synthetic)	88.6	3.95 + 0.07
Cascade (S2U + U2S, Lee et al., 2022) †	39.9	3.41 + 0.14

†: Not directly comparable because of differences on evaluation details.

Voice preservation

- Avoids potential misuse for creating “deepfakes”
  - No explicit speaker encoding
- Support speaker turns
  - Relies on the temporal synchronization between the synthesizer and the decoder
  - ConcatAug enables training on datasets without speaker turns
- Audio samples
  - Source (Spanish):
    - “En el fin de semana hemos hablado vino.” “Papá, mira, solo quiero saber de qué estás hablando.”
  - Prediction (English):
    - “on the weekend we talked about wine.” “dad look i just want to know what you're talking about.”

Thank you for listening!

Audio samples: