Using raw speech data for Speech-to-speech-translation

Sam Courtney

Background

Who I am:

- Music Engineering Tech grad student working on DSP, ML, and acoustics.
- Advised by Dr. Tom Collins, Dr Christopher Bennett

Why look into translational systems with IDSC?

- IDSC's mission of looking into cross disciplinary technologies.
- These technologies will connect us, speed us up on a societal scale.
- Translation requires understanding both the sound and the meaning of speech

What Problem I’m Investigating

- Speech-to-speech translation: mapping spoken audio in one language to spoken audio in another.
- Prosody: pitch/intonation/duration patterns that carry emotion and emphasis.
- Audio fidelity: preservation of quality, timbre (minimizing artifacts)
- ***most don’t tap into the seemingly “unlimited” amount of data: raw speech***
  - Scarcity of labeled parallel data + compute cost of better models.

Why This Matters

- Losing timbre/prosody makes speech sound robotic or identity-erased.
    - We’re not going to want to use these technologies for 8+ hours a day if they sound terrible.
- The lack of these technologies inherently builds walls between people

Key Idea

- Pretrain a direct S2ST model on massive amounts of raw, unlabeled speech using self-supervision, Then fine-tune on smaller labeled corpora. <- A great start!
- ***Potentially: investigate potential for continual improvement?***

Computational Challenge

- Working with a Million-hour dataset (Unsupervised People's Speech). (realistic!)
- Self-supervised learning = large batch sizes, long sequences.
- Direct S2ST = will likely involve acoustic encoder + linguistic latent encoder + vocoder; heavy GPU memory footprint.
- Need distributed training + multi-GPU pipeline.
- Need large-scale evaluation on quality metrics (PESQ, MOS, speaker similarity).

Thank you, IDSC!

- Multi-GPU large-batch training -> feasible
- Access high-capacity storage for multi-TB audio data
- Faster experiment cycles -> more comparisons/evaluations
- Mentorship on distributed training and data engineering