CMU's IWSLT 2023 Simultaneous Speech Translation System

Brian Yan, Jiatong Shi, Soumi Maiti, William Chen, Xinjian Li, Yifan Peng, Siddhant Arora and Shinji Watanabe

System Overview

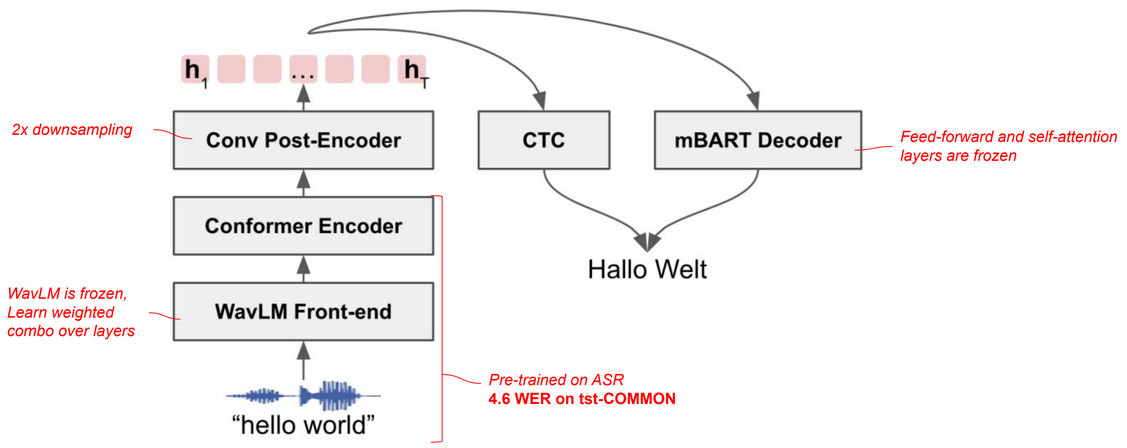
1. **Offline ST using joint CTC/attention** with self-supervised speech/text representations
2. **Offline-to-online adaptation** via chunk-based encoding and incremental beam search
3. **Cascaded simultaneous S2ST** by feeding incremental text outputs to a TTS model

Data:

- MuST-C + Tedlium (w/ MT pseudo-labels) for ST
- CommonVoice subset for TTS

Offline ST: Joint CTC/Attention + SSL/LLM

- Large scale model (~900M trainable params) with WavLM and mBART initializations



Cascaded Simultaneous S2ST

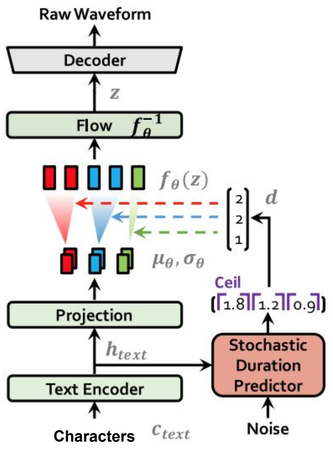


Image Source: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech

- Single-speaker VITS TTS model with character input
- CommonVoice data selection:
  - Evaluated the speech quality (DNSMOS) of top 5 most speakers by # of utterances
  - Set a threshold of 4.0 for selecting utterances
  - Choose the single speaker with most hours (12h)
- Cascaded inference on incremental text decodings (which correspond to ~2.5s of input speech)

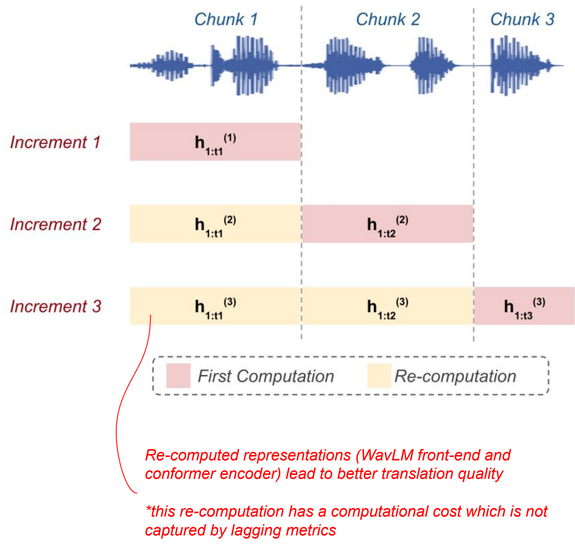
Offline-to-Online: Incremental Encoding/Decoding

- Directly use offline ST model for online inference via incremental strategies

Incremental Encoding

Re-compute encoder representations for each incremental chunk of speech

- 2s for speech-to-text
- 2.5s for speech-to-speech



Incremental Decoding

Algorithm 1 Beam search step with rewinding of unreliable hypotheses on non-final chunks and incremental pruning upon end-detection.

procedure BEAMSTEP(hyps, prevHyps, isFinal)
  newHyps = {}; endDetected = False
  for $y_{1:l-1} \in$ prtHs **do**
    attnCnds = top-k($P_{\text{Attn}}(y_l|X, y_{1:l-1}), \text{k} = p$)
    for $c \in$ attnCnds **do**
      $y_{1:l} = y_{1:l-1} \oplus \text{c}$
      $\alpha_{\text{CTC}} = \text{CTCScore}(y_{1:l}, X_{1:T})$
      $\alpha_{\text{Attn}} = \text{AttnScore}(y_{1:l}, X_{1:T})$
      $\beta = \text{LengthPen}(y_{1:l})$ *Joint CTC/attn scoring*
      $P_{\text{Beam}}(y_{1:l}|X) = \alpha_{\text{CTC}} + \alpha_{\text{Attn}} + \beta$
      newHyps$[y_{1:l}] = P_{\text{Beam}}(\cdot)$
      if (!isFinal) and (c is <eos> or repeat) **then**
        endDetected = True
        newHyps = prevHyps ▷ rewind
      else if l is max$L$ **then**
        endDetected = True *End-of-chunk detection*
      end if
    end for
  end for *Pruning*
  if endDetected **then** ▷ incremental pruning
    newHyps = top-k($P_{\text{Beam}}(\cdot), \text{k} = 1$)
  else ▷ standard pruning
    newHyps = top-k($P_{\text{Beam}}(\cdot), \text{k} = b$)
  end if
  return newHyps, endDetected
end procedure

Results

- 6% quality degradation from ST to SST (~60% lagging reduction)
- 12% quality degradation from SST to SS2ST (includes ASR errors from ASR-BLEU)

Model	Quality	Latency	
Offline Speech Translation (ST)	BLEU ↑	-	
Multi-Decoder CTC/Attn (Yan et al., 2023b)	30.1	-	-
WavLM-mBART CTC/Attn (Ours)	32.5	-	-
Simul Speech Translation (SST)	BLEU ↑	AL ↓	LAAL ↓
Time-Sync Blockwise CTC/Attn (Yan et al., 2023b)	26.6	1.93	1.98
WavLM-mBART CTC/Attn (Ours)	30.4	1.92	1.99
Simul Speech-to-Speech Translation (SS2T)	ASR-BLEU ↑	SO ↓	EO ↓
WavLM-mBART CTC/Attn + VITS (Ours)	26.7	2.33	5.67

-2.1 BLEU

-3.7 BLEU

Carnegie Mellon University
Language Technologies Institute
CMU-LTI WAVLab
{byan, jiatongs}@cs.cmu.edu