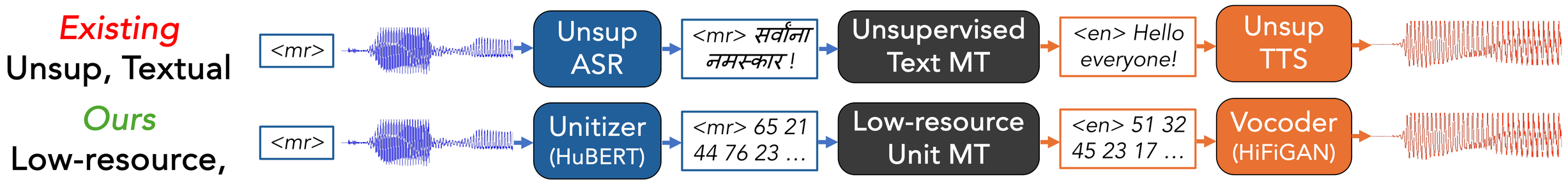
Textless Speech-to-Speech Translation With Limited Parallel Data

Anuj Diwan[1], *Anirudh Srinivasan*[1], *David Harwath*[1], *Eunsol Choi*[2]
[1] University of Texas at Austin [2] New York University

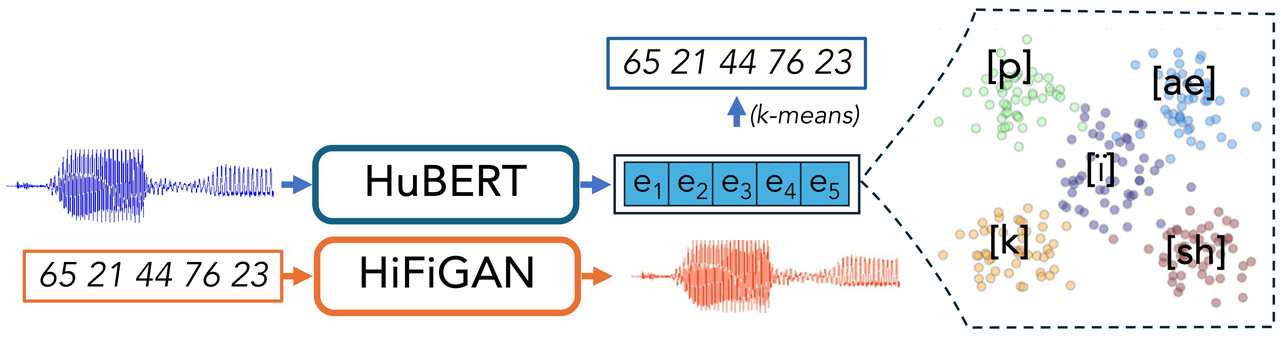
Website

Current S2ST models either rely on text as an intermediary or require extensive parallel speech data, limiting support for textless *and* low-resource languages. How can we bridge the gap?

We adapt an unsupervised text-based S2ST approach to the low-resource, textless S2ST setting by using **units** instead of text

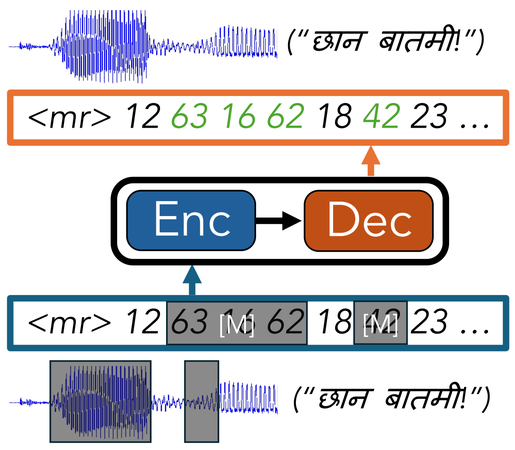


Step ⓪: Use existing unitizer and vocoder to convert speech ←→ units (self-supervised tokens with emergent phone-like structure)

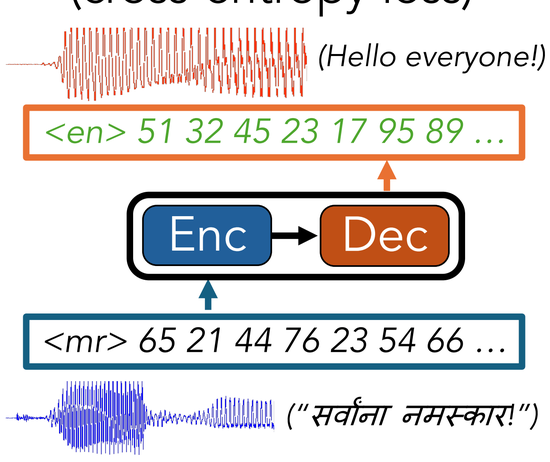


Steps ①-③: Train a low-resource unit Enc → Dec MT model (like mBART)

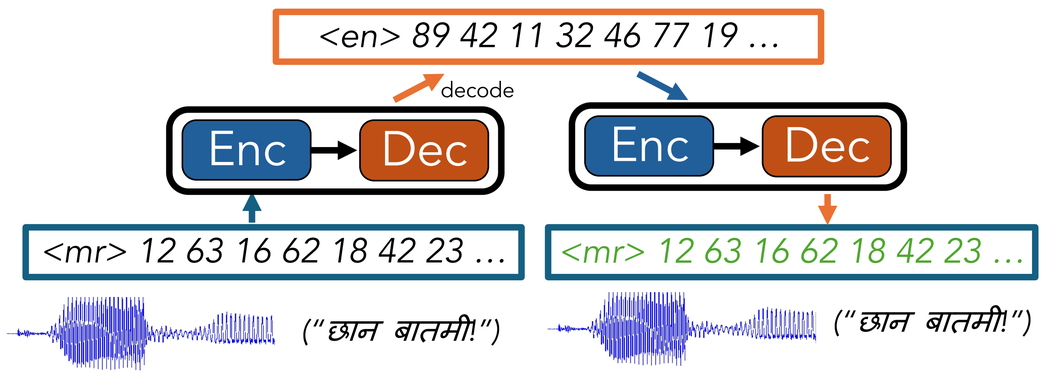
Step ①: Pretrain LM on monolingual per-lang data (masked denoising loss)



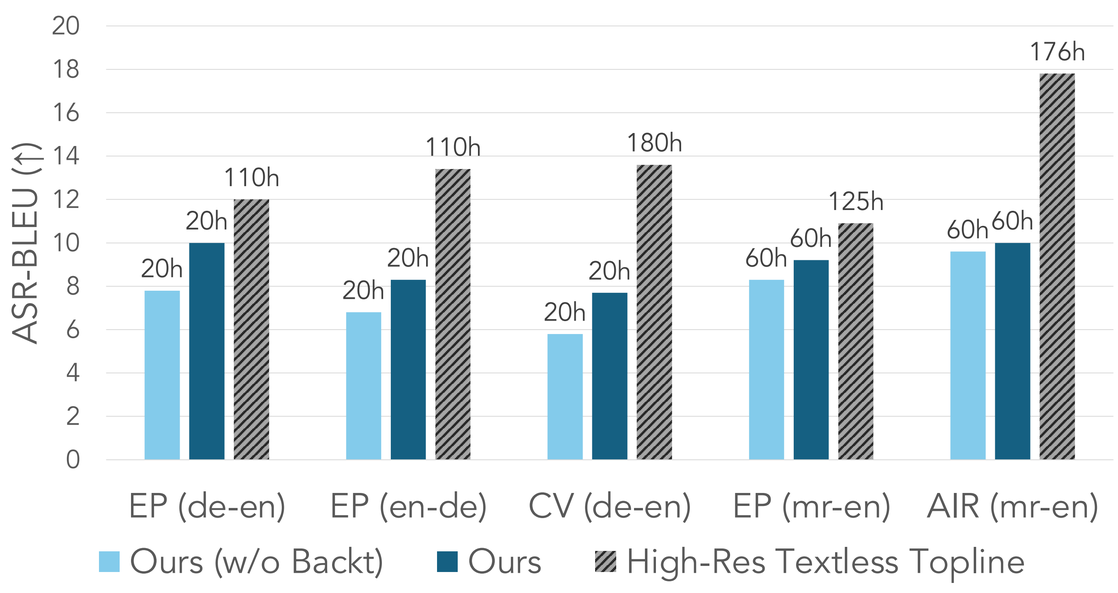
Step ②: Finetune LM on limited S2S translation data (cross-entropy loss)

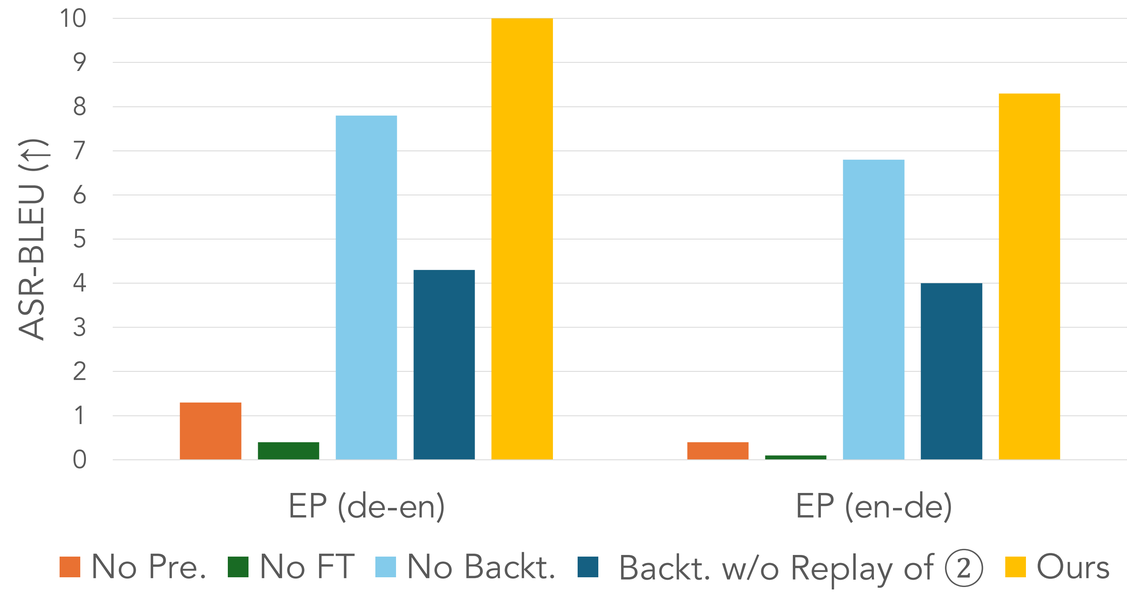


Step ③: Backtranslate LM on monolingual per-lang data (while replaying ②) (cross-entropy loss)



Experiments: *Languages*: en-de, en-mr. *Domains*: EP (Europarl), CV (Common Voice), AIR (All Ind Radio)





Main Takeaways

1. In some settings, our method is within 1-2 ASR-BLEU points of a high-res textless topline
2. Diff. domains exhibit diff. performance gaps

Ablations

1. In order of importance, FT > Pre. > Backt.
2. Replay is necessary for backtranslation to work

More analyses in the paper!

1. How to select an appropriate unitizer?
2. How do textless models compare to text-based models?
3. How does performance differ for short vs. long utterances?

Future Work

1. Scaling to stronger pretrained multilingual unit LMs, with potential for zero-resource textless S2ST
2. Using more semantic unitizers for efficient unit LM training