Enhancing End-to-End Conversational Speech Translation Through Target Language Context Utilization

Amir Hussein[1], Brian Yan[2], Antonios Anastasopoulos[3] , Shinji Watanabe[2] , Sanjeev Khudanpur[1]
[1]Johns Hopkins University, [2]Carnegie Mellon University, [3]George Mason University

Introduction

- End-to-end Speech Translation (E2E-ST): exciting advances BUT Translations from isolated utterances lack consistency.
- Context could help with ambiguity (pronouns, entities, homophones).
- Previous approaches **naively concatenate audio** as source language context [1].
- **Extended audio limitations:** memory limitation, hard to train.
- How to incorporate the context with minimum memory cost?
- How about additional contextual information, e.g., speaker ID?

Method

- **Proposed approach**: incorporate previous sentence translations as the initial condition for decoder.
- E2E-ST builds upon the **CTC/Attention** [2], decomposes the ST into ASR and translation.
- **SOC**: start of context; **SOS**: start of sentence; **EOS**: end of sentence

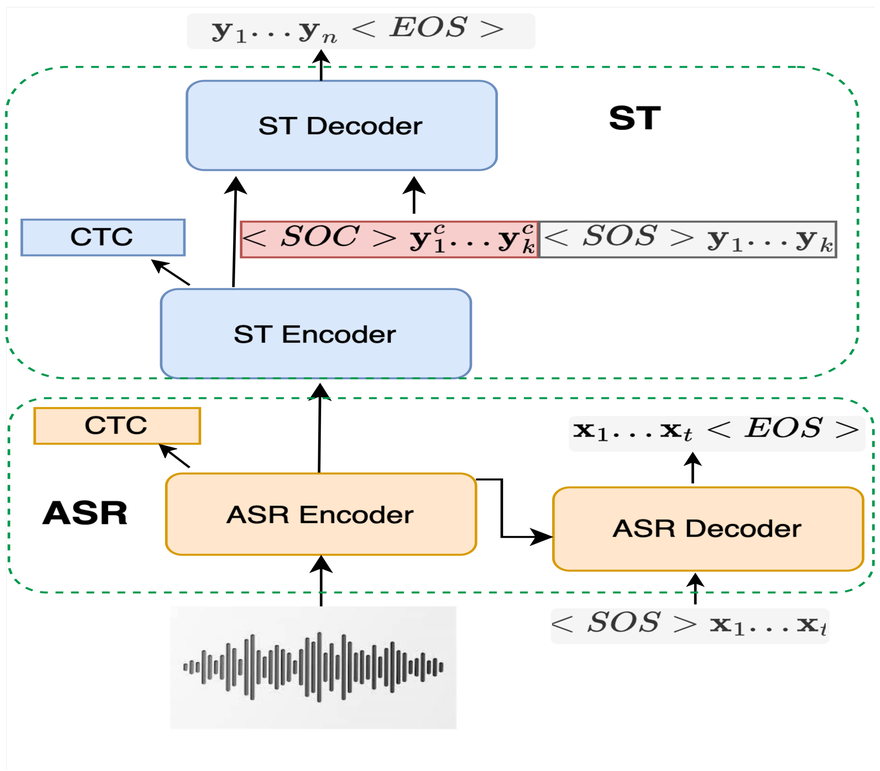


Figure 1. illustration of the proposed contextual E2E-ST approach

- Enrich the context with speaker role information:

[**Context**] [SpkA] I'm from Peru, and you? [SEP]
[SpkB] Puerto Rico.
[**Target**] [SpkA] Oh, from Puerto Rico, oh, ok.

Results

- Context bias: train with context and inference without context worse than baseline by up to -0.9 BLEU
- Context dropout: improves by up to +0.5 BLEU

ID	Train w/context	Decode w/context	Context Dropout	Fisher	CallHome	IWSLT22
1	✗	✗	-	29.8	25.9	19.7
2	✓	✓	-	**31.3**†	26.0	19.9
3	✓	✗	-	29.3	25.0	18.9
4	✓	✓	0.2	31.0†	**26.5**†	**20.2**†
5	✓	✗	0.2	30.1	25.8	19.8

-0.9, +0.5

Overall Improvements

- Incorporating gold context: up to +2.2 BLEU
- Exact decoding: previous predictions used as context for subsequent predictions.
- Multistage decoding: initial predictions from isolated utterances provide context for subsequent decoding stages. (+0.9 BLEU).

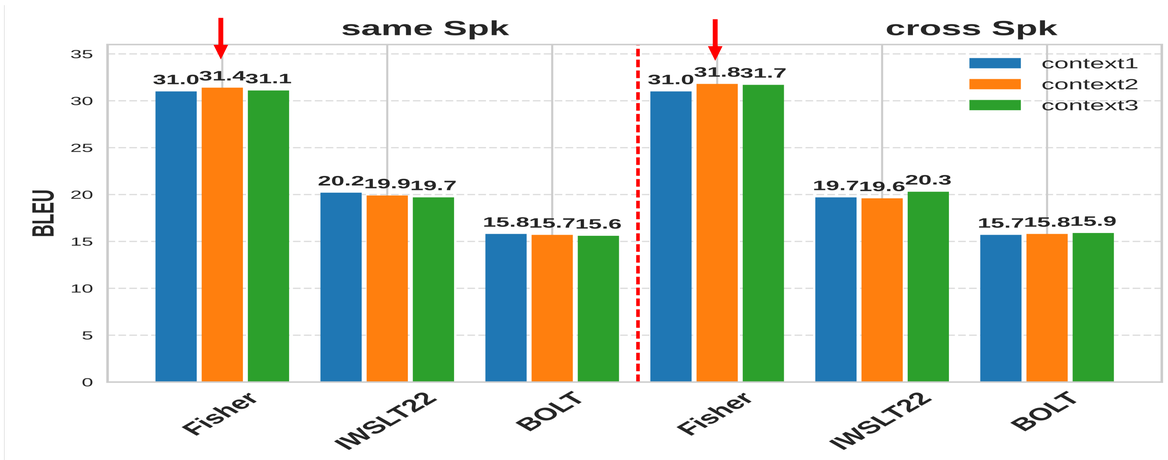
Controls context dependence (stages) and reduces error propagation.

Context type	Context size	Evaluation Sets Fisher	CallHome	IWSLT22	BOLT
Baseline	no-context	29.8	25.9	19.7	15.5
Gold	(2,2,3,3)	**31.8**†	**28.1**†	**20.3**†	**16.0**†
Hyp Exact	(2,2,3,3)	30.2†	25.9	19.8	15.6
Hyp Multistage	(2,2,3,3)	**30.7**†	**26.4**†	**19.8**	**15.9**†

+2.2, +0.9

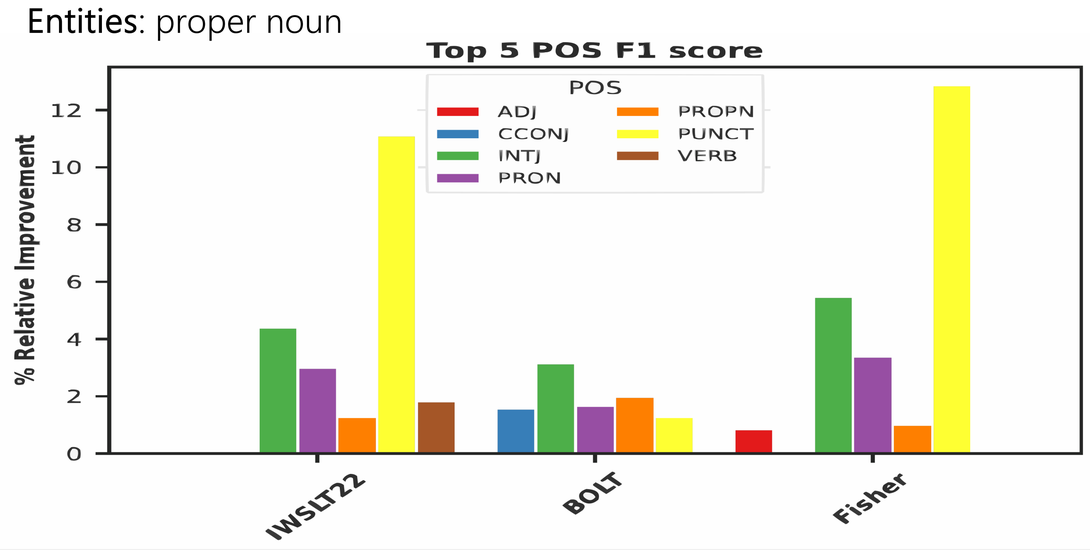
Context size and speaker role

- Cross speaker context outperforms same speaker context (+.4 BLEU)
- Optimal context size is between 2–3 utterances



Where do we improve?

- Highest relative improvements:
  1. **Style**: punctuations, interjections
  2. **Anaphora**: pronouns
  3. **Entities**: proper noun



Conclusion

- Incorporating context leads to significant improvements
- Some datasets benefit more from context (Spanish-English)
- Context dropout enhances robustness to context absence
- Speaker information further improves the performance
- Major improvements from context: context style, anaphora, entities

References

[1] B. Zhang et al., "Beyond sentence-level end-to-end speech translation: Context helps," in Proc. ACL, 2021, pp. 2566–2578
[2] B. Yan et al., "ESPnet-ST-v2: Multipurpose spoken language translation toolkit," in Proc. ACL, 2023, pp. 400–411.