ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446102

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
USA

ABSTRACT

Incorporating longer context has been shown to benefit machine translation, but the inclusion of context in end-to-end speech translation (E2E-ST) remains under-studied. To bridge this gap, we introduce target language context in E2E-ST, enhancing coherence and overcoming memory constraints of extended audio segments. Additionally, we propose context dropout to ensure robustness to the absence of context, and further improve performance by adding speaker information. Our proposed contextual E2E-ST outperforms the isolated utterance-based E2E-ST approach. Lastly, we demonstrate that in conversational speech, contextual information primarily contributes to capturing context style, as well as resolving anaphora and named entities.

Index Terms— Speech translation, contextual information, end-to-end models, conversational speech

1. INTRODUCTION

Speech translation (ST) is crucial for breaking language barriers and enhancing global communication. Seamlessly translating spoken conversations across languages impacts cross-cultural interactions, education, and diplomacy [1]. Traditionally, ST systems have been built by cascading automatic speech recognition (ASR) and machine translation (MT) models [2–8]. However, recently end-to-end speech translation (E2E-ST) approaches, in which the source speech is directly translated to the target language text, have gained more attention [9–13]. Despite the promising results of recent research advancements in E2E-ST systems, the produced translations often lack consistency in translation.

Consistency is crucial for language understanding, as the meaning of utterances often depends on the broader conversational context. In MT, source-side context is commonly utilized to address inaccurate choice of pronouns [14], mistranslations of ambiguous words [15], and general incoherence in translation [16]. Several MT studies showed that document level context [17–21] outperforms sentence level context [22–28]. Analogous to MT, incorporating contextual information in E2E-ST systems is expected to be valuable for coherent translation, resolving anaphora, and disambiguating words, especially homophones. To study the effectiveness of context on E2E-ST, researchers have used simple concatenation of audio input as context and the corresponding translation as target output, reporting improvements in pronoun and homophone translation [29]. However, utilizing source-side context comes with the challenge of encoding very long audio segments, which can easily lead to memory bottlenecks, especially with self-attention based networks [30].

To address this limitation, we propose a context-aware E2E-ST that leverages context in the output (target) side. In particular, in a conversational context, we incorporate the previously output sentences (in the target language) as the initial condition on the decoder side to generate the translation of the current input utterance. This enables the decoder to effectively utilize textual information from longer utterances and focus on vital parts of the previous context for more accurate and consistent translation. Unlike existing work, we focus on *conversational* speech translation, an essential facet of daily communication which presents unique challenges: 1) high context dependence for meaning, 2) informal and grammatically inconsistent language usage, and 3) data scarcity. Our study covers three language pairs: Tunisian Arabic-English, Spanish-English, and Chinese-English. Our contributions encompass: (i) a context-aware E2E-ST framework employing target language context, (ii) enhanced robustness to context absence through context dropout, and (iii) context enrichment with speaker information. As an additional contribution, we conduct an ablation study to assess the significance of each component (context size, context dropout, and speaker information) on final performance. Finally, we perform a part-of-speech-based analysis to identify where the primary improvements result from the incorporation of context.

2. CONTEXTUAL E2E-ST

In standard E2E-ST, the goal is to find the most probable target word sequence $\hat{\mathbf{Y}}$ of length N, out of all possible outputs $\mathbf{Y}^*$. This is done by selecting the sequence that maximizes the posterior likelihood $P(\mathbf{Y}|\mathbf{X})$, given a T-long sequence of D-dimensional speech features, represented as $\mathbf{X} = \{\mathbf{x}_t \in \mathbb{R}^D | t = 1, \cdots, T\}$. In our approach we incorporate K *previous* translations $\mathbf{Y}^{\text{cntx}} = \{\mathbf{y}_l \in \mathbb{V}^{\text{tgt}} | l = 1, \cdots, K\}$ in the target language $\mathbb{V}^{\text{tgt}}$. We incorporate the context as an initial condition on the decoder side. Therefore, our objective is to maximize the posterior likelihood given both the input speech and the context:

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}^*} \sum_{l=1}^{N} \log(P(\mathbf{Y}_l | \mathbf{Y}_{<l}, \mathbf{X}, \mathbf{Y}^{\text{cntx}})) \quad (1)$$

We enrich the context with speaker role information, encoded as speaker tags within $\mathbb{V}^{\text{tgt}}$. Figure 1 presents an illustrative example showcasing a target sentence along with a context of its two preceding sentences. This context is augmented with speaker role information: [SpkA] and [SpkB] represent to the first and second speaker in the conversation, respectively. The [SEP] tag indicates separation between the sentences within the context.

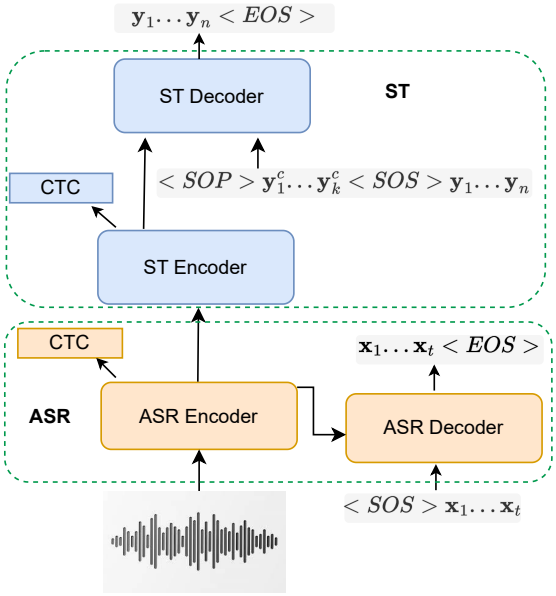

Fig. 1. Illustration of proposed contextual E2E-ST approach.

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

It's important to note that, unlike [31] where speaker labels are used for prediction, we employ them solely in the context as initial condition and do not predict them.

2.1. E2E-ST Architecture

Our proposed contextual E2E-ST builds upon the CTC/Attention architecture composed of conformer encoders with hierarchical CTC encoding and transformer decoders [13, 32] as depicted in Fig 1. The CTC/Attention approach decomposes ST into ASR and MT encoder-decoder models. The ASR encoder, $\mathrm{ENC}_{\mathrm{asr}}(\cdot)$, maps input speech $\mathbf{X}$ into hidden representation $\mathrm{H}^{\mathrm{E}}_{\mathrm{asr}}$ shown in Eq. (2).

$$H^{\mathrm{E}}_{\mathrm{asr}} = ENC_{\mathrm{asr}}(\mathbf{X}) \tag{2}$$

Following this, the representation $\mathrm{H}^{\mathrm{E}}_{\mathrm{asr}}$ from Eq. (2) serves as input to both the ST encoder, $\mathrm{ENC}_{\mathrm{st}}(\cdot)$, and the ASR decoder, $\mathrm{DEC}_{\mathrm{asr}}(\cdot)$, as demonstrated in Eq. (3,4)

$$H^{\mathrm{E}}_{\mathrm{st}} = ENC_{\mathrm{st}}(H^{E}_{\mathrm{asr}}) \tag{3}$$

$$H^{\mathrm{D}}_{\mathrm{asr}} = DEC_{\mathrm{asr}}(H^{E}_{\mathrm{asr}}) \tag{4}$$

Given $\mathrm{H}^{\mathrm{E}}_{\mathrm{st}}$ from Eq. (3), the ST encoder $\mathrm{DEC}_{\mathrm{st}}(\cdot)$ generates $\hat{\mathbf{y}}_l$ at each time step that is conditionally dependent on the hidden representation $\mathrm{H}^{\mathrm{E}}_{\mathrm{st}}$, previous target sequence $\mathbf{Y}_{1:t-1}$ and the context $\mathbf{Y}^{\mathrm{cntx}}$ of previous sentences shown in Eq. (5, 6).

$$H^{\mathrm{D}}_{\mathrm{st}} = DEC_{\mathrm{st}}(H^{E}_{\mathrm{st}}, \mathbf{Y}^{\mathrm{cntx}}_{1:K}, \mathbf{Y}_{1:l-1}) \tag{5}$$

$$P(\mathbf{y}_l|\mathbf{Y}_{1:l-1}, \mathbf{X}, \mathbf{Y}^{\mathrm{cntx}}) = \mathrm{Softmax}(H^{\mathrm{D}}_{\mathrm{st}}) \tag{6}$$

The unique characteristic of our approach lies in the use of context also during training. While other methods may attempt to predict the context, our model uses the context exclusively as an initial condition for the decoder, overcoming memory constraints of extended audio. During model training, the ST cross-entropy loss is computed solely for the target translation, excluding the context. The model is optimized with a multi-task learning objective [13] combining hybrid ASR attention $\mathcal{L}^{\mathrm{asr}}_{\mathrm{att}}$ and CTC $\mathcal{L}^{\mathrm{asr}}_{\mathrm{ctc}}$ as well as hybrid ST attention $\mathcal{L}^{\mathrm{st}}_{\mathrm{att}}$ and CTC $\mathcal{L}^{\mathrm{st}}_{\mathrm{ctc}}$ losses:

$$\begin{aligned}\mathcal{L} =& \alpha_3((1-\alpha_1)\mathcal{L}^{\mathrm{asr}}_{\mathrm{att}} + \alpha_1\mathcal{L}^{\mathrm{asr}}_{\mathrm{ctc}}) \\ &+ (1-\alpha_3)((1-\alpha_2)\mathcal{L}^{\mathrm{st}}_{\mathrm{att}} + \alpha_2\mathcal{L}^{\mathrm{st}}_{\mathrm{ctc}})\end{aligned} \tag{7}$$

where $\alpha's$ are used for interpolation.

3. EXPERIMENTS

3.1. Dataset

We demonstrate the efficacy of our proposed approach through evaluations on three conversational datasets (their statistics are summarized in Table 1): Fisher-CallHome Spanish English [33], IWSLT22 Tunisian Arabic-English [34], and BOLT Chinese-English [35]. These datasets contain 3-way data comprising telephone speech, source language transcriptions, and corresponding English translations. We use separated source and target vocabularies, each

Table 1. Statistics for the conversational ST corpora.

Corpus	Lang	#Hours		
		Train	Dev	Test
Spanish Fisher/Callhome	Sp-En	186.3	9.3	4.5/1.8
Tunisian IWSLT22	Ar-En	161.0	6.3	3.6
Chinese BOLT	Zh-En	110.6	8.5	8.5

consisting of 4K byte-pair-encoding [36] (BPE) units. All audios are resampled from 8kHz to 16kHz, augmented with speed perturbations (0.9, 1.0 and 1.1) and transformed into 83-dimensional feature frames (80 log-mel filterbank coefficients plus 3 pitch features). Additionally, we augment the features with `specaugment` [37], with mask parameters $(mT, mF, T, F) = (5, 2, 27, 0.05)$ and bi-cubic time-warping. During scoring we report the results in terms of case-sensitive BLEU with punctuation. We also measure the statistical significance of improvement using paired bootstrap resampling with sacreBLEU[38].

3.2. Baseline configuration

We conduct all experiments by customizing the ESPnet toolkit [32]. For the encoders $\mathrm{ENC}_{\mathrm{asr}}$ in Eq. (2) and $\mathrm{ENC}_{\mathrm{st}}$ in Eq. (3), we employ the conformer architecture [39] consisting of 12 blocks for $\mathrm{ENC}_{\mathrm{asr}}$ and 6 blocks for $\mathrm{ENC}_{\mathrm{st}}$. Both encoders are configured with 2048 feed-forward dimensions, 256 attention dimensions, and 4 attention heads. The transformer architecture is employed for $\mathrm{DEC}_{\mathrm{asr}}$ and $\mathrm{DEC}_{\mathrm{st}}$ in Eq. (4), each with 6 decoder blocks and the same configuration as the encoders. The model has 72M parameters. We follow ST best practices [32], we first pretrain the ASR module followed by fine-tuning of the entire E2E-ST model for the translation task. Our training configuration remains consistent for both pretraining and fine-tuning, employing Adam optimizer with a learning rate of 0.001, warmup-steps of 25K, a dropout-rate of 0.1 and 40 epochs. We use joint training with hybrid CTC/attention by setting CTC weight (α_1, α_2) in Eq. (7) to 0.3 and the weight that combines ASR and ST losses

Table 2. Comparison of the BLEU scores according to contextual information quality using one previous utterrance as context. †: denotes a statistically significant difference ($p < 0.01$) compared to the no-context baseline.

Context Type	Fisher	CallHome	IWSLT22
No-context	29.8	25.9	19.7
Random	29.9	25.2	19.5
Gold	**31.3**†	**26.0**	**19.9**

(α_3) to 0.3. During inference, we use a beam size of 10 and length penalty of 0.3.

3.3. Contextual E2E-ST configuration

In the Contextual E2E-ST model (described in §2), we retain the same configuration as the baseline. The only difference is that during training with teacher-forcing, the decoder initial condition is the preceding sentences alongside a start-of-sentence token. Heuristically, we limit any long contextual sentences to the last 50 tokens. Upon visual inspection, we found that these truncated sentences effectively capture the contextually relevant information. The previous sentence considered a part of context only if they are from the same recording.[1] We will refer to (Gold) context when employing ground-truth translations and to (Hyp) when utilizing the model's predictions. During training we only use the Gold context, however during inference we explore both Gold and Hyp context, simulating both an oracle and a more realistic scenario.

4. RESULTS

4.1. Contextual ST results

We examined the effect of preceding context on model performance by comparing Gold previous context with the no-context baseline and randomly selected sentences unrelated to the ground truth. The results, shown in Table 2, use a context size of one preceding sentence. The analysis indicates that, compared to the no-context baseline, gold context improves performance, with a maximum BLEU increase of $+1.5$. Random context, on the other hand, lowers performance up to a maximum BLEU reduction of -0.7. This outcome affirms that incorporating even a single sentence as context yields improvements, which stem exactly from high-quality context rather than other artifacts.

4.2. Context dropout

Next we explore the bias resulting from training with gold context when no context is available during inference (results in Table 3). A model trained with context but applied to inference without context (third row) shows a degradation of almost -1 BLEU points even compared to the non-contextual baseline (first row). This clearly demonstrates the context-trained model's strong inclination to depend on context during inference. To overcome this limitation we propose to use context dropout: during training, target-side context is probabilistically included or not. We experiment with various percentages of context dropout ([0.2–0.7]) and find that 0.2 yields the best results across all datasets. Now, using context dropout (row four) improves the BLEU score by up to $+0.5$ during inference with available context, compared to scenarios without it (row two). But more importantly, row five shows that the model trained with context dropout exhibits robustness to inference without context, even slightly surpassing the context-less baseline's performance.

[1]The first utterance of each conversation has no associated context.

Table 3. Comparison of the BLEU scores according to the bias towards contextual information using context size of one. † denotes a statistically significant difference ($p < 0.01$) compared to the no-context baseline.

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

4.3. Context size and speaker role

In this part we investigate the impact of context size and speaker information. We explore previous gold context of size [1,2,3], which may include utterances from a different speaker (cross Spk) or we can select context only from the same speaker. Results are in Fig 2. From Fig 2, we conclude that cross-speaker context consistently outperforms same-speaker context, evident in the BLEU scores showing improvements of $+.4$, $+0.1$, and $+0.2$ for Fisher, IWSLT22, and BOLT respectively. The optimal context size for cross speaker is between 2–3: 2 for Fisher, 3 for IWSLT22 and BOLT.

4.4. Using model-predicted context

Now we turn to using model-predicted context (Hyp) instead of Gold context, using the optimal configuration from §(4.2,4.3) with 0.2 context dropout and cross-speaker context. We examine two decoding approaches: a) **Exact decoding:** at every step, the model's previous predictions are used as the contextual input for subsequent predictions. This method may exhibit error propagation because initial predictions can affect later ones. b) **Multi-stage decoding:** initial predictions from isolated utterances provide context for subsequent predictions. This method controls context dependence and computational cost by adjusting the number of decoding stages. Initial predictions for every stage can be made independently, enabling parallelization, however

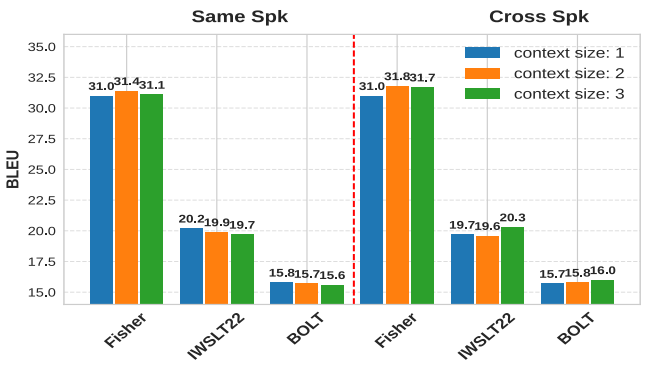

Fig. 2. Comparison of the BLEU scores according to context size and speaker information. Cross-speaker context outperforms same-speaker context, but the optimal context size varies (2–3).

Source	Context	Ref	Hyp-w/context	Hyp-no/context
para que me dejen tranquila no hablo inglés ahí	I know who it is and say, no thank you, no thank you. oh, sometimes I pretend I don't speak English	so that they leave me alone I don't speak English, there	because it's really calm, I don't speak English, there is	Why do you feel calm? No, a month, from there
¿pero es americana o?	we laugh a little and she tells me about her family	But is she American or?	But is she American or?	But is he American or?
و الله تبدا قاعده مسكينه	Did you see by God By God I sometimes go out at four from here and figure out she's still working	By God she's staying poor her	By God she's staying	I swear you'll be sitting
قعدت ايه	Maybe she gained weight after giving birth	She stayed yes	She stayed yes	She's staying yes

Fig. 3. Examples of translations with and without context. Color code: {**Green**: matching with reference; **Red**: incorrect; **Orange**: not matching with reference but similar meaning; **Blue**: related contextual tokens}

Table 4. Comparison of using predicted Hyp context (**Exact** and **Multistage**), and best Gold context results. **Context Size** column indicates the number of previous utterances used as context for each respective dataset. †: denotes a statistically significant difference ($p < 0.01$) compared to the no-context baseline.

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**†

the computational cost becomes multiplied by the number of stages compared to the baseline. Additionally, we provide a comparative analysis against the best results achieved using gold context, shown in Table 4. The overall improvement of using Gold context, compared to the baseline, is up to +2.2 BLEU points. On the other hand, it is noteworthy that Hyp Multistage context outperforms Hyp Exact. For multistage decoding, we use only one stage as we noticed a degradation after the first stage, which supports our error propagation hypothesis. The overall improvement using Multistage approach compared to the no-context baseline is up to +0.9 BLEU. Additionally, on the BOLT dataset, the Multistage approach nearly matches the results achieved with Gold context.

4.5. Where do we improve?

While BLEU scores provide an aggregate measure over the entire test corpus, they may fail to capture the nuances that a context-informed model can address. Even if the difference in BLEU scores is minor, our model retains critical contextual information. To better understand where our model improves, we analyze the part-of-speech (POS) tags of the predictions compared to the reference, using a state-of-the-art transformer-based POS-tagger from Spacy.[2] We focus on Spanish (Fisher), IWSLT22, and BOLT. We compute the relative improvement in F1 score for each POS tag and visualize the top five improvements in Fig 4. We observe a consistent pattern where the highest relative improvement is from *Context Style* (PUNCT, INTJ) followed by *Anaphora* (PRON), and *Entities* (PROPN). It is worth noting that a large portion of BOLT dataset is not punctuated, which contributes to the relatively lower PUNCT improvement. Examples highlighting the benefits of contextual information are presented in Figure 3. In these examples, context helps in disambiguating: a) homophones: *para que* → *so that* vs *por que* → *why*, b) pronouns: *americana o* vs *americano* → *she* vs *he is American* c) coherence in style: *By God* vs *swear* d) verb tense: *stayed* vs *staying*.

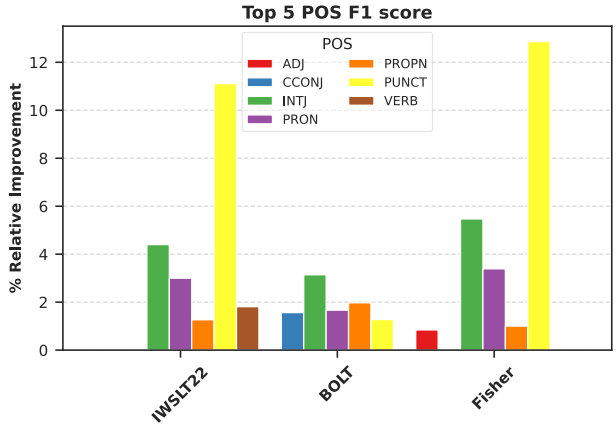

Fig. 4. Top 5 relative improvements in F1 score across different Part-of-Speech Tags.

5. CONCLUSION

We developed an end-to-end contextual Speech Translation (ST) model that leverages target language context, significantly outperforming a no-context baseline across three conversational speech translation datasets. This highlights the pivotal role of high-quality contextual information, including cross-speaker information, in enhancing model performance. Despite context-trained models exhibiting strong contextual dependence during inference, this can be effectively mitigated by implementing context dropout. Moreover, the comparison between different decoding strategies, using model-predicted context, showed that a multi-stage decoding approach provides significant improvements and reduces the risk of error propagation. Our analysis demonstrates that contextual information contributes primarily to context style, anaphora, and entities.

6. ACKNOWLEDGEMENTS

This work was carried out during the 2023 Summer Camp for Applied Language Exploration at Human Language Technology Center of Excellence. In addition, this work was partially supported by NSF CCRI Grant No 2120435.

[2]https://spacy.io/models/en

7. REFERENCES

[1] O. Köksal and N. Yürük, "The role of translator in intercultural communication," *International Journal of Curriculum and Instruction*, vol. 12, no. 1, pp. 327–338, 2020.

[2] F. Stentiford and M. Steer, "Machine translation of speech," in *Speech and language processing*, 1990, pp. 183–196.

[3] E. Ansari *et al.*, "FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN," in *Proceedings of the 17th International Conference on Spoken Language Translation*, 2020, pp. 1–34.

[4] A. Waibel *et al.*, "Janus: A speech-to-speech translation system using connectionist and symbolic processing strategies," in *Acoustics, speech, and signal processing, IEEE international conference on*, 1991, pp. 793–796.

[5] N. Bertoldi and M. Federico, "A new decoder for spoken language translation based on confusion networks," in *Proc. ASRU*, 2005, pp. 86–91.

[6] M. Sperber, J. Niehues, and A. Waibel, "Toward robust neural machine translation for noisy input sequences," in *Proceedings of the 14th International Conference on Spoken Language Translation*, 2017, pp. 90–96.

[7] J. Pino *et al.*, "Harnessing indirect training data for end-to-end automatic speech translation: Tricks of the trade," in *Proceedings of the 16th International Conference on Spoken Language Translation*, 2019.

[8] J. Yang *et al.*, "JHU IWSLT 2022 dialect speech translation system description," in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, 2022, pp. 319–326.

[9] A. Berard *et al.*, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *Proceedings of the NIPS Workshop on end-to-end learning for speech and audio processing*, 2016.

[10] A. Berard *et al.*, "End-to-end automatic speech translation of audiobooks," in *Proc. ICASSP*, 2018, pp. 6224–6228.

[11] S. Dalmia *et al.*, "Searchable hidden intermediates for end-to-end models of decomposable sequence tasks," in *Proc. NAACL*, 2021, pp. 1882–1896.

[12] M. Gaido *et al.*, "End-to-end speech-translation with knowledge distillation: FBK@IWSLT2020," in *Proceedings of the 17th International Conference on Spoken Language Translation*, 2020, pp. 80–88.

[13] B. Yan *et al.*, "ESPnet-ST-v2: Multipurpose spoken language translation toolkit," in *Proc. ACL*, 2023, pp. 400–411.

[14] L. K. Guillou, "Incorporating pronoun function into statistical machine translation," 2016.

[15] A. Rios Gonzales, L. Mascarell, and R. Sennrich, "Improving word sense disambiguation in neural machine translation with sense embeddings," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 11–19.

[16] E. Voita, R. Sennrich, and I. Titov, "When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion," in *Proc. ACL*, 2019, pp. 1198–1212.

[17] L. Wang *et al.*, "Exploiting cross-sentence context for neural machine translation," in *Proc. EMNLP*, 2017, pp. 2826–2831.

[18] L. Miculicich *et al.*, "Document-level neural machine translation with hierarchical attention networks," in *Proc. EMNLP*, 2018, pp. 2947–2954.

[19] E. Voita *et al.*, "Context-aware neural machine translation learns anaphora resolution," in *Proc. ACL*, 2018, pp. 1264–1274.

[20] A. Lopes *et al.*, "Document-level neural MT: A systematic comparison," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 2020, pp. 225–234.

[21] Z. Zheng *et al.*, "Towards making the most of context in neural machine translation," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, Ed., 2020, pp. 3983–3989.

[22] S. Shon *et al.*, "Context-aware fine-tuning of self-supervised speech models," in *Proc. ICASSP*, 2023, pp. 1–5.

[23] T. Hori *et al.*, "Advanced long-context end-to-end speech recognition using context-expanded transformers," in *Proc. Interspeech*, H. Hermansky *et al.*, Eds., 2021, pp. 2097–2101.

[24] S. Kim and F. Metze, "Acoustic-to-word models with conversational context information," in *Proc. ACL*, 2019, pp. 2766–2771.

[25] S. Kim, S. Dalmia, and F. Metze, "Cross-attention end-to-end ASR for two-party conversations," in *Proc. Interspeech*, G. Kubin and Z. Kacic, Eds., 2019, pp. 4380–4384.

[26] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," *ArXiv preprint*, vol. abs/2212.04356, 2022.

[27] Z. Tian *et al.*, "How to make context more useful? an empirical study on context-aware neural conversational models," in *Proc. ACL*, 2017, pp. 231–236.

[28] B. Gain, R. Haque, and A. Ekbal, "Not all contexts are important: The impact of effective context in conversational neural machine translation," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[29] B. Zhang *et al.*, "Beyond sentence-level end-to-end speech translation: Context helps," in *Proc. ACL*, 2021, pp. 2566–2578.

[30] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, I. Guyon *et al.*, Eds., 2017, pp. 5998–6008.

[31] N. Kanda *et al.*, "Serialized output training for end-to-end overlapped speech recognition," in *Proc. Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds., 2020, pp. 2797–2801.

[32] B. Yan *et al.*, "Espnet-st-v2: Multipurpose spoken language translation toolkit," *ArXiv preprint*, vol. abs/2304.04596, 2023.

[33] M. Post *et al.*, "Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus," in *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, 2013.

[34] E. Ansari *et al.*, "FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN," in *Proceedings of the 17th International Conference on Spoken Language Translation*, 2020, pp. 1–34.

[35] Z. Song *et al.*, "Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 1699–1704.

[36] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. EMNLP*, 2018, pp. 66–71.

[37] D. S. Park *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, G. Kubin and Z. Kacic, Eds., 2019, pp. 2613–2617.

[38] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186–191.

[39] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds., 2020, pp. 5036–5040.