Improving End-to-End Speech Translation by Imitation-Based Knowledge Distillation with Synthetic Transcripts

Rebekka Hubert[⋆], Artem Sokolov[⋆,†], Stefan Riezler[⋆]

Heidelberg University[⋆] & Google Research[⋆]

{hubert,sokolov,riezler}@cl.uni-heidelberg.de

Imitation Learning for automatic speech translation (AST)

- Large text-based NMT expert model continues and corrects translations starting from an AST input prefix.
- AST student model imitates continuations of NMT expert.

- **Advantages:**
  - **Knowledge transfer** from large general-domain text translation models to speech translation.
  - AST student model can **explore its own output space** during training.
  - NMT expert's corrections of student's output enrich training data with **examples of successful recovery from errors**.
  - Theoretical guarantees showing **prediction error scales at most linearly with time** for imitation learning, unlike quadratic scaling for standard teacher forcing.

Related Work

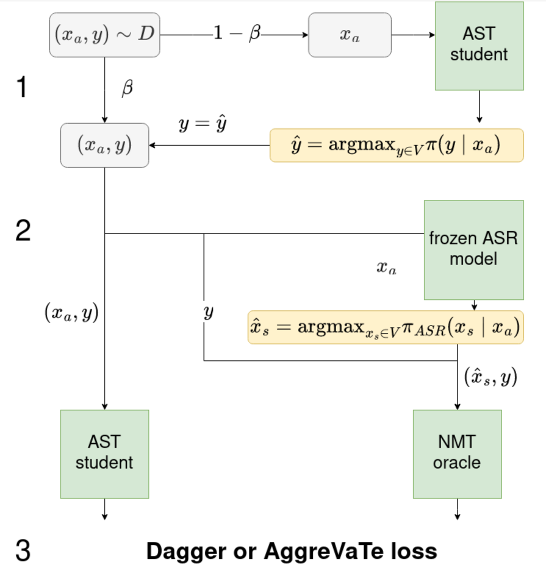
- Knowledge distillation for AST [Liu et al., 2019a, Gaido et al., 2020]
- Imitation Learning for text-based NMT [Lin et al., 2020, Hormann and Sokolov, 2021]
- More standard tricks of the trade: Data augmentation using text-to-speech and translations of manual transcripts [Jia et al., 2019, Pino et al., 2019, Pino et al., 2020], incorporated via multi-task learning [Weiss et al., 2017, Anastasopoulos and Chiang, 2018]

- **Shortcomings:** Most related approaches still rely on in-domain **manual source transcripts** for data augmentation or to train teacher model.

Our work: Imitation Learning in AST **without manual transcripts**

- Input synthetic source transcript from ASR model to NMT expert.
- Initialize encoder of AST student with weights from corresponding ASR model.

- **Advantages:**
  - Incorporate knowledge from same pre-trained ASR model into NMT expert and AST student, without access to manual source transcripts (e.g., WHISPER [Radford et al., 2022], AudioGPT [Huang et al., 2023], AudioPaLM [Rubenstein et al., 2023]).
  - This work: Proof-of-concept experiment using open-source ASR system but discarding transcripts.
  - **Key question:** Is the NMT expert still able to function as **error-correcting oracle** when faced with synthetic source transcripts?



$(x_a, y) \sim D$
$1-\beta$
x_a
AST student
1
β
$y = \hat{y}$
(x_a, y)
$\hat{y} = \mathrm{argmax}_{y \in V} \pi(y \mid x_a)$
2
frozen ASR model
x_a
(x_a, y)
y
$\hat{x}_s = \mathrm{argmax}_{x_s \in V} \pi_{ASR}(x_s \mid x_a)$
$(\hat{x}_s, y)$
AST student
NMT oracle
3
Dagger or AggreVaTe loss

Adaptation of DAgger (Dataset Aggregation) [Ross et al., 2011] to AST

- **NMT expert:** Predict next-step correction v_t^* given source synthetic transcript x_s and partial AST student hypothesis $y_{<t}$:

$$v_t^* = \operatorname*{argmax}_{v \in V} \pi^*(v \mid y_{<t}; x_s).$$

- **AST student:** Minimize negative log-likelihood of student AST model π given speech input x_a w.r.t. optimal expert prediction v_t^*:

$$\mathcal{L} = \mathbb{E}[-\sum_{t=1}^{T} \log \pi(v_t^* \mid y_{<t}; x_a)],$$

OR: Minimize cross-entropy w.r.t. expert model's distribution π^*:

$$\mathcal{L} = \mathbb{E}[-\sum_{v \in V} \pi^*(v \mid y_{<t}; x_s) \cdot \log \pi(v \mid y_{<t}; x_a)].$$

Data and Models

- English-German AST on CoVoST2 [Wang et al., 2021] (430 hours) and MuST-C [Di Gangi et al., 2019] datasets (408 hours)
- NMT expert: Transformer from Facebook's submission to WMT19 [Ng et al., 2019], based on Big Transformer architecture [Vaswani et al., 2017], trained for text-based translation.
- AST student: RNNs and Base Transformers, based on fairseq framework [Ott et al., 2019, Wang et al., 2020], trained for speech translation.
- ASR helper: Base Transformers, based on fairseq framework [Ott et al., 2019, Wang et al., 2020], trained on manual transcripts.

Variant	Expert Input	Loss
Standard	-	CE
KD$^+$ [Liu et al., 2019b]	gold	CE
SynthKD$^+$	synthetic	CE
IKD [Lin et al., 2020]	gold	$\mathcal{L}_{\text{IKD}}$
IKD$^+$ [Lin et al., 2020]	gold	$\mathcal{L}_{\text{IKD}^+}$
SynthIKD (ours)	synthetic	$\mathcal{L}_{\text{IKD}}$
SynthIKD$^+$ (ours)	synthetic	$\mathcal{L}_{\text{IKD}^+}$

- Variants of Knowledge Distillation (KD) and Imitation Learning (IKD) indicated by $^+$ access expert's full probability distribution instead of expert's optimal action.
- Prefix Synth- indicates use of synthetic source transcripts.

Model	Hypotheses	#	Decoding Setup	Transcripts	dev-BLEU↑
RNN	full	1	AST	-	11.9
		2	ASR transcribes, NMT expert translates	-	21.8
	partial	3	AST starts, NMT expert completes	gold	21.9
		4	AST starts, NMT expert completes	synthetic	15.6
Transformer	full	5	AST	-	16.7
		6	ASR transcribes, NMT expert translates	-	25.4
	partial	7	AST starts, NMT expert completes	gold	25.4
		8	AST starts, NMT expert completes	synthetic	19.9

- Question: Does NMT expert have higher quality than AST student?
- Lower bound by student (#1, #5); upper bound by expert (#2, #6).
- Expert completion improves over student (#3, #7), even with synthetic source input (#4, #8).

Achitecture		Models	CoVoST2 dev	CoVoST2 test	MuST-C dev	MuST-C test
RNN	baseline	Standard	13.6	10.0	14.6	14.1
	baseline	KD$^+$	**14.6**	**11.1**	**17.9**	**17.2**
	baseline	IKD$^+$	13.1	10.1	15.7	14.9
	ours	SynthKD$^+$	14.1	10.6	16.9	15.9
	ours	SynthIKD$^+$	12.8	9.7	16.3	15.1
Transformer	baseline	Standard	18.4	14.2	19.5	19.4
	baseline	KD$^+$	21.3	17.7	17.7	22.2
	baseline	IKD$^+$	**21.8**	18.4	23.2	23.3
	ours	SynthKD$^+$	21.7	18.0	22.5	22.6
	ours	SynthIKD$^+$	**21.8**	**18.5**	**23.5**	**23.5**

- Statistically insignificant differences between IKD with synthetic or gold source transcripts.
- IKD outperforms KD and standard baselines for Transformers.
- Smaller gains for RNNs because of their lower translation quality.

AggreVaTe (Aggregate Values to Imitate) [Ross and Bagnell, 2014]

- Expert produces full continuation $y^*_{>t}$ till the end:

$$y^*_{>t} = \underset{y_{>t}}{\operatorname{argmax}}\, \pi^*(y_{>t} \mid y_{<t} + a_t; x).$$

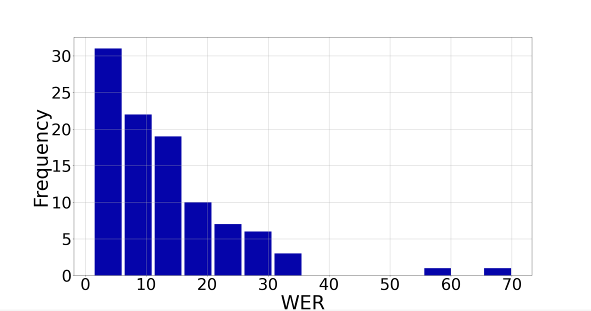
  where $y_{\geq t}$ is a continuation, a_t is an exploratory next-step prediction, and argmax is implemented as beam search.

- Student trained to decrease square loss between logit Q of selected action a_t and BLEU of predicted suffix $y^*_{>t}$:

$$\mathcal{L} = \mathbb{E}[\sum_{t=1}^{T} \big(\sigma(Q(a_t \mid y_{<t}; x)) - \text{BLEU}(y^*_{>t})\big)^2].$$

IL Algorithm	Model	Data	CoVoST2 BLEU↑ dev	CoVoST2 BLEU↑ test	CoVoST2 TER↓ dev	CoVoST2 TER↓ test	MuST-C BLEU↑ dev	MuST-C BLEU↑ test	MuST-C TER↓ dev	MuST-C TER↓ test
Dagger	Standard	gold	18.4	14.2	69.1	77.1	19.5	19.4	70.8	69.4
	IKD^{+}	gold	21.8	18.4	63.7	70.0	23.2	23.3	67.4	65.6
	SynthIKD^{+}	synth	21.8	**18.5**	63.6	69.8	**23.5**	23.5	67.2	65.6
	Warm-start Model	**Data**	**BLEU↑ dev**	**BLEU↑ test**	**TER↓ dev**	**TER↓ test**	**BLEU↑ dev**	**BLEU↑ test**	**TER↓ dev**	**TER↓ test**
			sentence-BLEU reward-to-go							
AggreVaTe	Standard	gold	18.7	14.6	68.2	76.0	19.9	19.9	70.2	68.1
	Standard	synth	18.7	14.6	68.2	75.9	20.0	19.7	70.1	68.7
	IKD^{+}	gold	**22.1**	**18.5**	**63.1**	69.6	**23.5**	23.4	67.4	65.7
	SynthIKD^{+}	synth	**22.1**	**18.5**	**63.1**	69.7	**23.5**	**23.6**	**67.0**	65.6
			TER reward-to-go							
	Standard	gold	18.7	14.7	67.8	75.4	20.0	19.9	70.0	68.5
	Standard	synth	18.7	14.6	67.9	75.6	19.9	19.6	69.8	68.4
	IKD^{+}	gold	22.0	**18.5**	**63.1**	**69.4**	23.3	23.4	67.3	65.5
	SynthIKD^{+}	synth	**22.1**	**18.5**	**63.1**	69.6	**23.5**	**23.6**	**67.0**	**65.3**

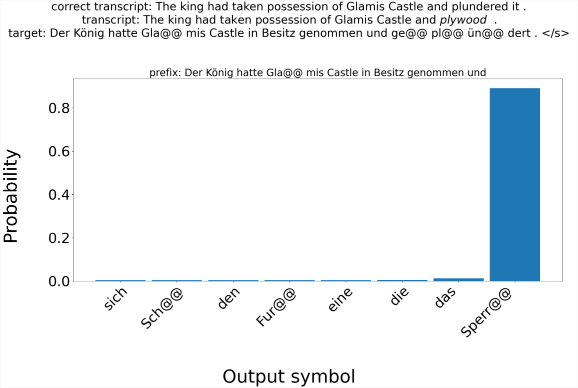
- No significant difference between training with sentence-BLEU reward-to-go or TER.
- No improvement over DAgger — one-step corrections sufficient!



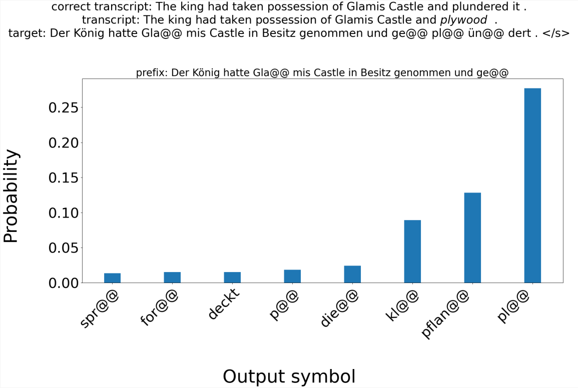
- Sentence-wise WER of ASR Transformer on 100 error samples from CoVoST2:
  - Mostly single-digit number of errors.

#	**Error Type**	**Freq**
1	omitted tokens	2
2	surface form error	17
3	contentual error, correct target in top-1	5
4	contentual error, correct target in top-8	12
5	critical error, expert predicts correctly due to pre-fix	32
6	critical error, expert does not predict correctly	32

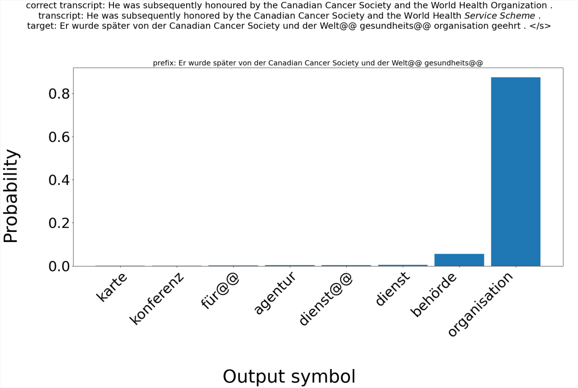
- ASR errors in lines 1-4 do not hinder NMT expert to produce correct output token.
- **ASR errors in line 5 can be corrected with help of prefix.**



- ASR: "The king had taken possession of Glamis Castle and plywood."
- NMT prefix: "Der König hatte Glamis Castle in Besitz genommen und "



- ASR: "The king had taken possession of Glamis Castle and plywood."
- NMT prefix: "Der König hatte Glamis Castle in Besitz genommen und ge"
- **Fluency preferred over correct translation of wrong source transcript.**



- ASR: “World Health Service Scheme”
- NMT prefix: “Er wurde später von der Canadian Cancer Society und der Weltgesundheits”
- **Expert prefers correct proper name due to context knowledge.**

Key findings for imitation learning with synthetic transcripts:

- Requires **large NMT expert** since error correction driven by **language modeling capability** of expert.
- Student needs access to **full expert distribution** (not just best expert prediction as in original DAgger).
- **Single-step corrections sufficient**, thus greedy search applicable.
- **Straightforward extension** to pre-trained ASR models using proprietary data.
- **Don't be afraid of error propagation in imitation learning!**

Thank you!

Anastasopoulos, A. and Chiang, D. (2018).
Tied multitask learning for neural speech translation.
In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, New Orleans, Louisiana.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (2019).
MuST-C: a Multilingual Speech Translation Corpus.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota.

Gaido, M., Di Gangi, M. A., Negri, M., and Turchi, M. (2020).
End-to-end speech-translation with knowledge distillation: FBK@IWSLT2020.
In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*, Online.

Hormann, L. and Sokolov, A. (2021).
Fixing exposure bias with imitation learning needs powerful oracles.
CoRR, abs/2109.04114.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J.-B., Liu, J., Ren, Y., Zhao, Z., and Watanabe, S. (2023).
AudioGPT: Understanding and generating speech, music, sound, and talking head.
CoRR, abs/2304.12995.

Jia, Y., Johnson, M., Macherey, W., Weiss, R. J., Cao, Y., Chiu, C.-C., Ari, N., Laurenzo, S., and Wu, Y. (2019).

Leveraging weakly supervised data to improve end-to-end speech-to-text translation.
In *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK.

Lin, A., Wohlwend, J., Chen, H., and Lei, T. (2020).
Autoregressive knowledge distillation through imitation learning.
In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Liu, Y., Xiong, H., Zhang, J., He, Z., Wu, H., Wang, H., and Zong, C. (2019a).
End-to-End Speech Translation with Knowledge Distillation.
In *Proceedings of INTERSPEECH*, Graz, Austria.

Liu, Z., Sun, M., Zhou, T., Huang, G., and Darrell, T. (2019b).
Rethinking the value of network pruning.
In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019).
Facebook FAIR's WMT19 news translation task submission.
In *Proceedings of the Fourth Conference on Machine Translation (WMT)*, Florence, Italy.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019).
fairseq: A fast, extensible toolkit for sequence modeling.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, Minnesota.

Pino, J. M., Puzon, L., Gu, J., Ma, X., McCarthy, A. D., and Gopinath, D. (2019).

Harnessing indirect training data for end-to-end automatic speech translation: Tricks of the trade.
Hong Kong, China.

Pino, J. M., Xu, Q., Ma, X., Dousti, M. J., and Tang, Y. (2020).
Self-training for end-to-end speech translation.
In *Proceedings of INTERSPEECH*, Shanghai, China.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022).
Robust speech recognition via large-scale weak supervision.
CoRR, abs/2212.04356.

Ross, S. and Bagnell, J. A. (2014).
Reinforcement and imitation learning via interactive no-regret learning.
CoRR, abs/1406.5979.

Ross, S., Gordon, G. J., and Bagnell, J. A. (2011).
A reduction of imitation learning and structured prediction to no-regret online learning.
In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA.

Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borsos, Z., de Chaumont Quitry, F., Chen, P., Badawy, D. E., Han, W., Kharitonov, E., Muckenhirn, H., Padfield, D. R., Qin, J., Rozenberg, D., Sainath, T. N., Schalkwyk, J., Sharifi, M., Tadmor, M. D., Ramanovich, Tagliasacchi, M., Tudor, A., Velimirovi'c, M., Vincent, D., Yu, J., Wang, Y., Zayats, V., Zeghidour, N., Zhang, Y., Zhang, Z., Zilka, L., and Frank, C. H. (2023).
AudioPaLM: A large language model that can speak and listen.

CoRR, abs/2306.12925.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017).
Attention is all you need.
In *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., and Pino, J. (2020).
Fairseq S2T: Fast speech-to-text modeling with fairseq.
In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, Suzhou, China.

Wang, C., Wu, A., Gu, J., and Pino, J. (2021).
CoVoST 2 and Massively Multilingual Speech Translation.
In *Proceedings of INTERSPEECH*, Brno, Czechia.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (2017).
Sequence-to-sequence models can directly translate foreign speech.
In *Interspeech*, Stockholm, Sweden.