Multilingual Speech Translation with Efficient Finetuning of Pretrained Models

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino,
Alexei Baevski, Alexis Conneau, Michael Auli
Facebook AI

UBC-NLP RG
May 5th
Peter Sullivan

Overview

- Motivation
- Additional Background Context
- Model
- Experiments
- Results

Motivation

- Speech Translation (ST) doesn’t have enough data for End-to-End (E2E) training in many languages.
- Cascade models dominate, but with clear downsides (error propagation)
- Unlabeled Pre-training + Transfer Learning might solve the data scarcity issues for E2E
- But fine-tuning large Acoustic and LMs needs to be efficient

Additional Background Context

Stoian et al. 2020, Bansal et al. 2019

- Low Resource ST relies on a pre-trained Encoder from High Resource ASR
- Language of pre-training and amount of data doesn’t matter so much as getting a decent WER of the ASR module

Background cont.

Liu et al. 2020 - mBART

- Adapt self-supervised training to multilingual MT through denoising pre-training. (Similar to Lewis et al. 2019 but multilingual)

Baevski et al. 2020 - wav2vec 2.0

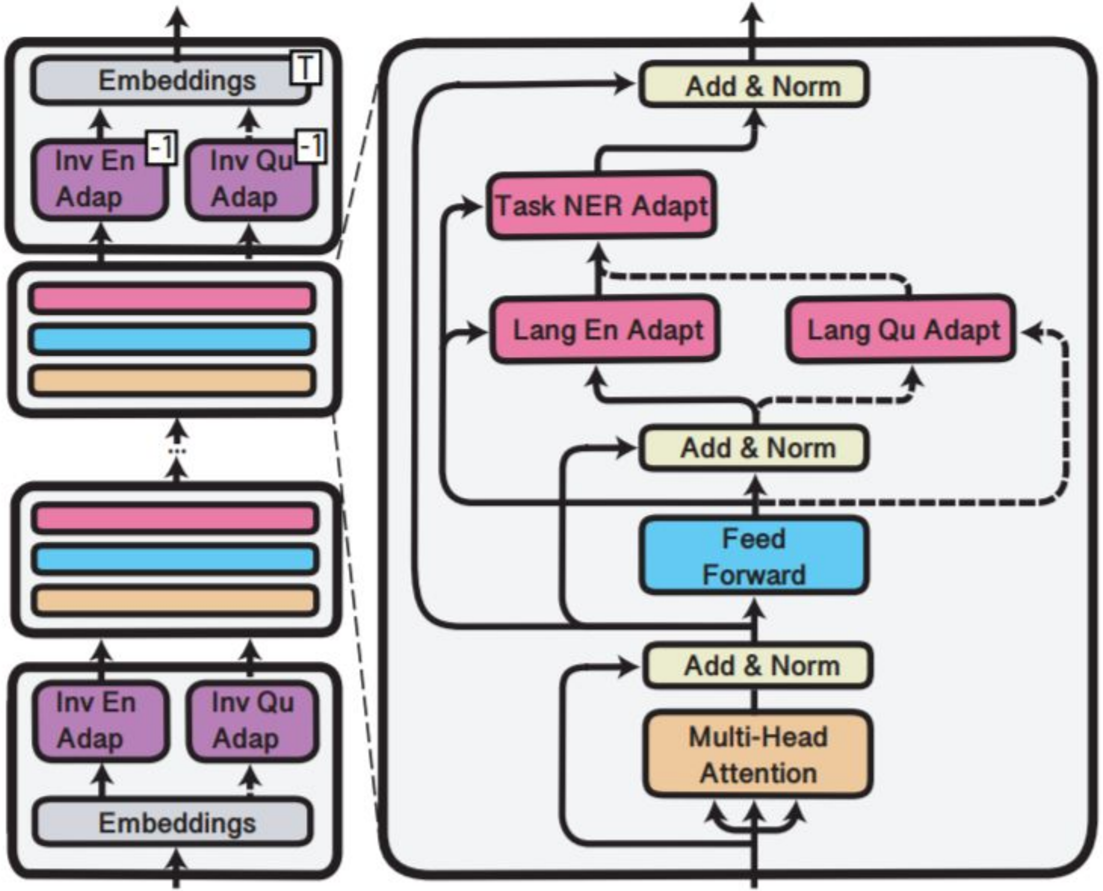
- Latest iteration of wav2Vec framework, add Transformer context network in addition to contrastive loss from wav2vec and quantization layers (vq-wav2vec)

Background cont.

Pfeiffer et al. 2020

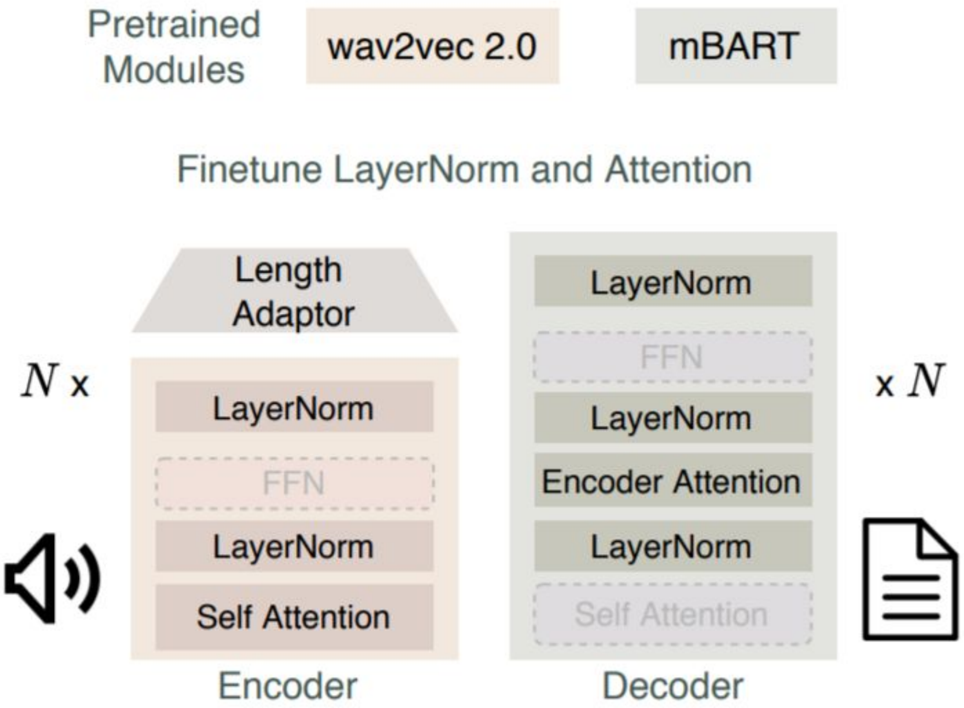
- Low Resource tasks with mBERT / XLM-R etc. suffer from lack of model capacity on unseen data
- Adapter modules can be added to solve this

See also Houlsby et al. 2019



This Paper - XMEF (CrossModal Efficient Finetuning):

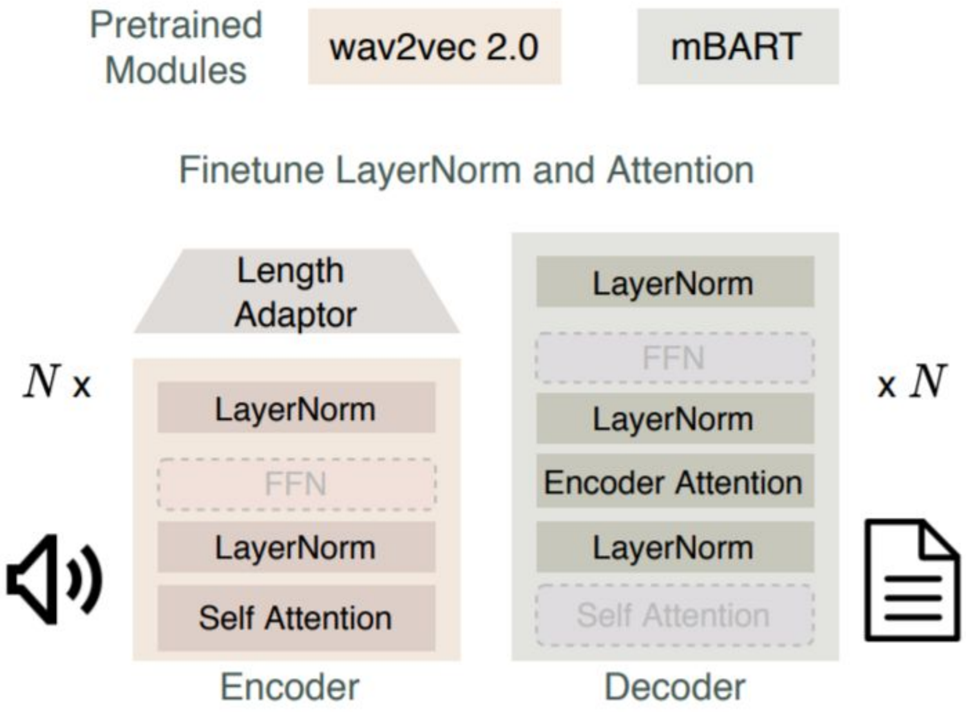
- Use pretrained Encoder + Decoder
- Only fine-tune Layer Norm and Attention (LNA)
- Joint train on Speech+Text
- Zero shot transfer
- Many-to-Many translation without parallel data



LNA fine-tuning

Intuition:

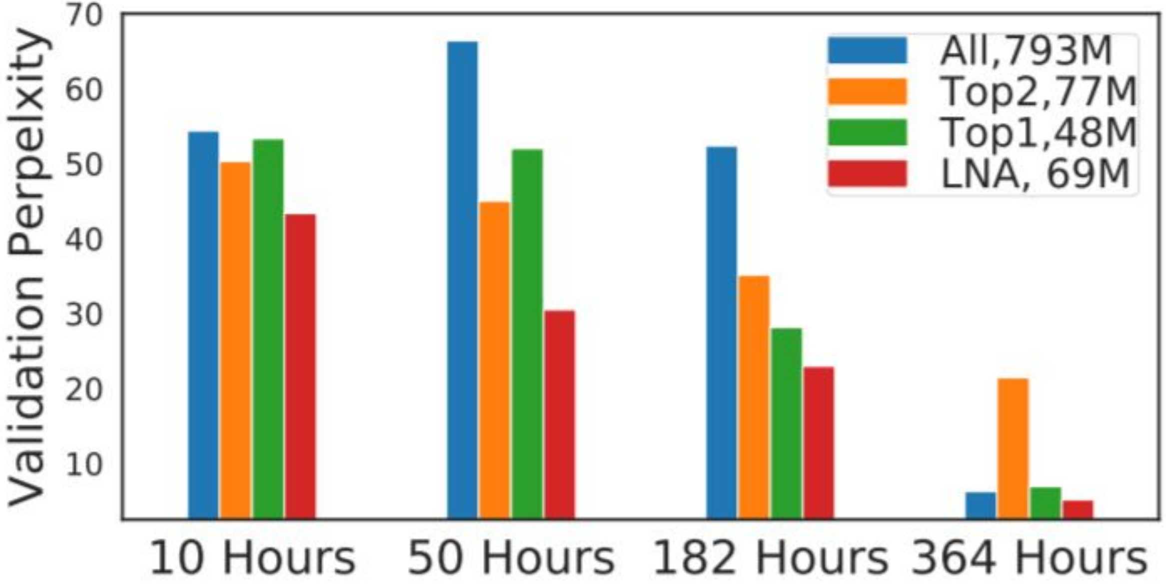
- Layer Norm (LN) originally trained on pre-training statistics
- Encoder Attention of MT, trained on Text-to-Text not speech
- Self-Attention might aid in learning multilingual structure



LNA vs. full fine-tuning

En-De Dev results from CoVoST 2

Hours indicate amount of data used in training



LNA-Min vs Encoder Self-Attention

De-En Dev results from CoVoST 2

Min = Only FT Layer Norm and Encoder Attention

ESA = Min w/Encoder Self-attention

Input = Feature extractor

LNA-Ablation

En-De Dev results from CoVoST 2

Enc	Dec	PPL ↓	Params (%)
LN	LN + EA	5.17	69.4M (8.8%)
- LN	- LN	37.66	69.3M (8.7%)
	- EA	5.97	19.0M (2.4%)
	+ SA	5.26	119.8M (15.1%)
+ SA		5.53	170.2M (21.5%)

Experiments

Datasets:

- CoVoST 2 (Wang et al. 2020)
  - En->X and X-> En covering many languages, many with <10 or <4 hrs.
- Europarl ST (Iranzo-Sanchez et al. 2020)
  - Large Parallel Data from European Parliament (En, De, Es, Fr, It, Pt)

Findings - Zero Shot Speech Side (CoVoST 2)

				Train					**Zero-shot**
	Enc	Dec	Params.	Fr	De	Es	Ca	It	Pt
LNA-E,D	LN+SA	LN+EA	170.7M	**32.4**	**24.9**	**31.6**	**28.6**	**24.0**	**8.2**
LNA-D	All	LN+EA	384.8M	31.6	23.7	31.0	27.8	23.2	7.6
Finetune All	All	All	793.0M	27.1	17.7	27.8	21.7	18.9	5.1
ASRPT+Multi				23.1	15.3	21.2	19.9	14.9	4.4
Supervised (Multi) SOTA (Wang et al., 2020b)				26.5	17.6	27.0	23.1	18.5	6.3

- Train on 5 Ls -> En, test on PT -> En (BLEU)

Findings - Zero Shot Text Side (CoVoST 2)

				Train				Zero-shot
	Enc	Dec	Params.	De	Fa	Tr	Zh	Ja
LNA-E,D	LN	LN+EA	69.4M	22.1	17.7	13.4	29.2	22.9
LNA-E,D	LN+SA	LN+EA	170.7M	23.8	19.2	14.2	30.6	29.2
LNA-D	All	LN+EA	384.8M	**24.9**	**19.8**	15.2	**32.7**	**30.6**
LNA-E	LN+SA	All	477.6M	22.0	18.1	14.2	29.5	0.8
Finetune All	All	All	793.0M	24.1	19.6	**15.6**	32.4	0.4
ASRPT+Multi				9.5	10.9	6.8	23.5	0.0
Supervised (Multi) SOTA (Wang et al., 2020b)				17.3	14.5	10.7	28.2	31.9

- Train on En -> 4Ls, test on En -> Ja (BLEU)

Findings - Select CoVoST 2 results (European)

	High Resource				Low Resource					
→ **En**	Fr	De	Es	Ca	It	Ru	Pt	Nl	Sl	Sv
Train Hours	264	184	113	136	44	18	10	7	2	2
Scratch-BL	24.3	8.4	12.0	14.4	0.2	1.2	0.5	0.3	0.3	0.2
+ ASR PT	26.3	17.1	23.0	18.8	11.3	14.8	6.1	3.0	3.0	2.7
+ Multi.	26.5	17.5	27.0	23.1	18.5	4.7	6.3	5.0	0.7	0.5
+mBART	28.1	19.7	28.1	24.0	19.9	2.7	6.2	8.1	0.5	1.4
LNA-E,D (170.7M)	**33.8***	**26.7***	**34.0***	**29.5***	**26.1***	**21.1**	**19.2**	**14.1***	**4.6**	<u>**5.9**</u>
LNA-D (384.8M)	<u>**35.0***</u>	**28.2***	<u>**35.2***</u>	<u>**31.1***</u>	**27.6***	**22.8**	<u>**24.1***</u>	**14.2***	**5.0**	**5.0**
Finetune All (793.0M)	**33.0***	**24.5***	**33.6***	**28.0***	**25.2***	**20.2**	**19.5**	**9.4**	**4.6**	**4.8**
Joint Training (1.05B)	**33.5***	**28.6***	**33.5***	**30.6***	**26.6***	**17.6**	**12.0**	<u>**15.0***</u>	**3.9**	2.6
+ Extra MT Data	**34.4***	<u>**29.6***</u>	**34.4***	**30.6***	<u>**27.7***</u>	<u>**27.7***</u>	**14.6**	**14.5***	<u>**5.2**</u>	**3.4**
Prev. E2E SOTA	27.0	18.9	28.0	24.0	11.3	14.8	6.1	8.4	3.0	2.7
Cascade SOTA	29.1	23.2	31.1	27.2	22.9	25.0	22.7	10.4	7.0	11.9

Findings - Select CoVoST 2 results (Low Resource/Dist.)

→ **En**	Fa	Zh	Tr	Et	Mn	Ar	Lv	Cy	Ta	Ja	Id	Avg.
Train Hours	49	10	4	3	3	2	2	2	2	1	1	
ASR (WER)	62.4	45.0	51.2	65.7	65.2	63.3	51.8	72.8	80.8	77.1	63.2	
Baseline	1.9	1.4	0.7	0.1	0.1	0.3	0.1	0.3	0.3	0.3	0.4	
+ ASR PT	3.7	5.8	3.6	0.1	0.2	4.3	2.5	2.7	0.3	1.5	2.5	
+ Multi.	2.4	5.9	2.3	0.6	0.1	0.4	0.6	1.9	0.1	0.1	0.3	7.0
+ mBART	3.3	5.4	2.4	0.7	0.2	0.5	0.6	1.4	0.1	0.2	0.2	7.3
LNA-E,D (170.7M)	**4.0**	**6.2**	<u>**5.5**</u>	**1.3**	<u>**1.0**</u>	3.7	**4.6**	2.8	**0.7**	**1.7**	**2.9**	12.5
LNA-D (384.8M)	3.6	**6.0**	**4.8**	<u>**1.5**</u>	**0.9**	2.8	<u>**4.9**</u>	2.3	<u>**0.8**</u>	**1.7**	<u>**3.7**</u>	12.6
Finetune All (793.0M)	3.7	<u>**6.5**</u>	**4.0**	**1.4**	**1.0**	3.3	**4.9**	2.1	**0.5**	<u>**2.1**</u>	**3.4**	11.2
Joint Training (1.05B)	<u>**6.1***</u>	5.4	3.3	0.7	0.2	0.8	**2.7**	1.0	0.1	0.3	0.5	10.7
+ Extra MT Data	**5.0**	**6.2**	**4.0**	0.8	0.3	1.0	**3.6**	1.1	0.2	0.5	0.5	11.7
Prev. SOTA	3.7	5.9	3.7	0.9	0.2	4.3	2.5	3.3	0.3	1.5	2.5	
Cascade	5.8	11.4	9.3	3.8	1.0	12.3	7.2	7.4	0.4	3.8	11.8	

Findings - CoVoST 2 En Speech

En →	Ar	Ca	Cy	De	Et	Fa	Id	Ja
Scratch-BL	8.7	20.2	22.2	13.6	11.1	11.5	18.9	26.9
+ ASR PT	12.1	21.8	23.9	16.5	13.4	13.5	20.8	29.6
+ Multi.	13.0	22.3	23.7	17.3	13.9	14.5	20.3	31.9
LNA-E,D-BL (69.4M)	12.0	18.8	12.9	**20.3***	15.0	**15.9***	**24.4***	31.4
LNA-E,D (69.4M)	**15.3***	20.3	13.2	**23.2***	**18.6***	**19.6***	**26.5***	**36.9***
LNA-E,D (170.7M)	**17.4***	22.2	14.8	**25.3***	**21.0***	**20.1***	**27.6***	**38.4***
LNA-E (477.6M)	**17.2***	**29.5***	**30.3***	**25.2***	**20.7***	**19.8***	**28.5***	**37.8***
Finetune All (793.0M)	**17.7***	**30.1***	**30.0***	**25.2***	**21.1***	**20.3***	**28.9***	**38.1***
Joint Training (1.05B)	**18.0***	**30.9***	**30.6***	**25.8***	**22.1***	**21.5***	**29.9***	**39.3***
Prev. E2E SOTA	13.9	23.6	25.1	18.4	15.1	15.5	22.0	33.0
Cascade SOTA	14.3	25.0	25.6	19.4	15.4	14.1	23.1	33.8

En →	Lv	Mn	Sl	Sv	Ta	Tr	Zh	Avg.
Scratch-BL	11.5	6.6	11.5	20.1	9.9	8.9	20.6	
+ ASR PT	13.1	9.2	16.1	22.3	11.2	10.2	25.7	
+ Multi.	14.1	10.2	17.1	22.3	11.7	10.7	28.2	18.1
LNA-E,D-BL (69.4M)	14.3	6.9	17.9	**26.1***	12.6	10.8	21.8	
LNA-E,D (69.4M)	**17.9***	**12.0***	**21.1***	**27.5***	**14.6***	**14.1***	**32.1***	20.9
LNA-E,D (170.7M)	**20.1***	**13.3***	**23.0***	**29.6***	**16.4***	**15.5***	**33.0***	22.5
LNA-E (477.6M)	**20.2***	**14.1***	**23.5***	**30.0***	**16.8***	**16.2***	**32.8***	24.2
Finetune All (793.0M)	**20.8***	**14.1***	**23.6***	**30.4***	**17.1***	**16.3***	**33.7***	24.5
Joint Training (1.05B)	**21.5***	**14.8***	**25.1***	**29.6***	**17.8***	**17.0**	**33.3**	25.1
Prev. E2E SOTA	15.2	11.0	18.3	24.1	12.8	11.7	31.3	
Cascade SOTA	15.6	11.7	18.9	24.8	13.7	11.7	26.9	

Findings - Europarl ST results Zero Shot

		Target					
		De	En	Es	Fr	It	Pt
Source	De		12.8/**20.6**	10.2/**13.8**	11.6/**14.9**	6.6/**8.6**	10.4/**13.0**
	En	13.1/**22.5***		23.1/**32.3***	22.1/**30.0***	14.9/**21.5**	20.7/**28.4**
	Es	9.2/**12.1**	18.9/**26.0**		19.0/**21.8**	13.3/**15.4**	20.0/**21.9**
	Fr	9.8/**13.6**	19.8/**27.9***	18.6/**21.7**		13.8/**15.2**	19.7/**21.4**
	It	10.1/**11.9**	19.8/**25.6**	18.8/**20.8**	19.1/**20.0***		19.8/19.2
	Pt	9.0/**11.4**	19.0/**24.1**	19.8/19.6	18.1/**18.6**	15.6/**16.1**	

- Shaded - Supervised Directions (En -> X or X -> En)
- All others are Zero Shot

Discussion

- XMEF proves effective at adapting pretrained models to new unseen languages
- Improvement over Cascade SOTA on many languages is a significant achievement, but does not hold for Low Resource X -> En
- Not a direct comparison to Adapter works (shortcoming)
- Future work might be in automatically learning layers to fine-tune (Guo et al. 2019)

References

- Stoian, M. C., Bansal, S., & Goldwater, S. (2020, May). Analyzing ASR pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7909-7913). IEEE.
- Li, X., Wang, C., Tang, Y., Tran, C., Tang, Y., Pino, J., ... & Auli, M. (2020). Multilingual Speech Translation with Efficient Finetuning of Pretrained Models. *arXiv preprint arXiv:2010.12829*.
- Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2020). Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning* (pp. 2790-2799). PMLR.
- Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., & Feris, R. (2019). Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4805-4814).

References

- Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726-742.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Wang, C., Wu, A., & Pino, J. (2020). Covost 2: A massively multilingual speech-to-text translation corpus. *arXiv preprint arXiv:2007.10310*.
- Iranzo-Sánchez, J., Silvestre-Cerdà, J. A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., ... & Juan, A. (2020, May). Europarl-ST: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8229-8233). IEEE.
- Bansal, S., Kamper, H., Livescu, K., Lopez, A., & Goldwater, S. (2018). Low-resource speech-to-text translation. *arXiv preprint arXiv:1803.09164*.