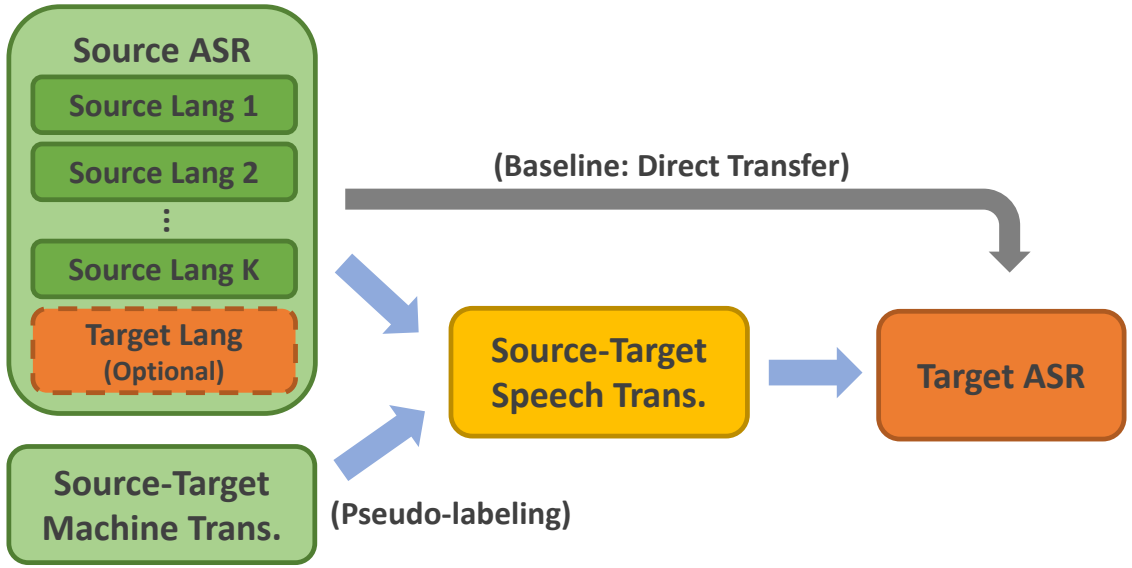
Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation

Changhan Wang, Juan Pino, Jiatao Gu

INTERSPEECH 2020

FACEBOOK

- Cross-lingual transfer from high-resource ASR to low-resource ASR
- Attention-based encoder-decoder models
- Transfer via speech-to-text translation (ST) task, so that decoder is pre-trained in the target language
- Up to 24.6% WER reduction on Common Voice and IARPA Babel compared to direct transfer



End-to-End Model Architecture

- Attention-based encoder-decoder models
- Same model architecture for source/target ASR and ST tasks
- Experiments on BLSTM-based encoder/decoder
- Similar application to Transformer-based models
- [Future work] extension to transducer models

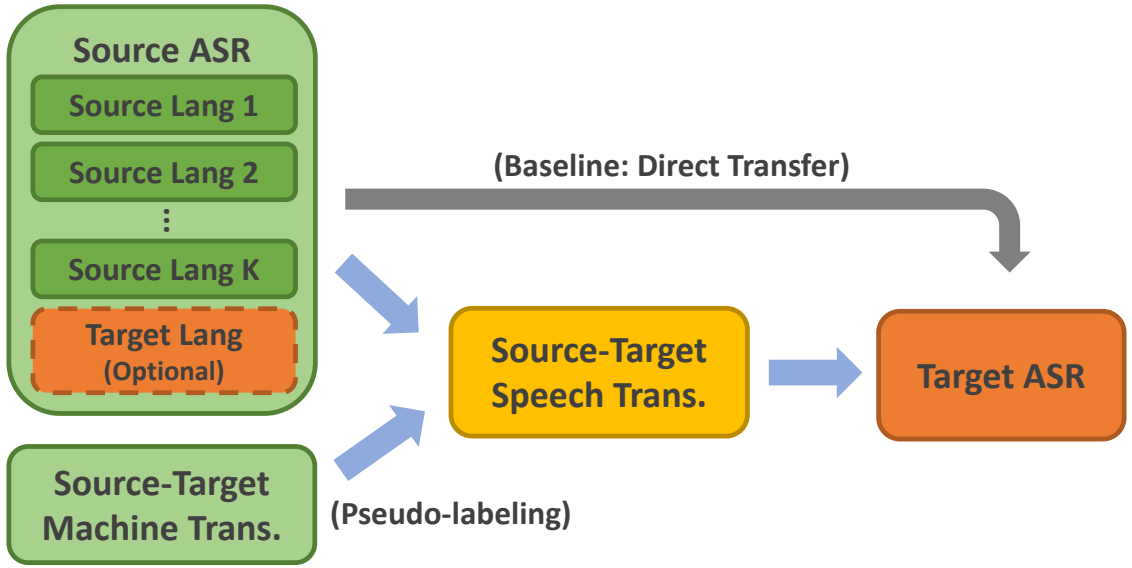
Speech Translation Trained with Pseudo-Labels

- Human labels for ST are expensive
- Pseudo-label source ASR data into target language with machine translation (MT) models
- Sequence-level knowledge distillation (KD) of MT labels
- KD-ed labels are simplified and easier to train on

Pre-training ASR on Speech Translation

- Direct transfer (baseline): source ASR → target ASR
- ST-based transfer:
  - [Basic] Source ASR → ST → target ASR
  - [Simplified] ST → target ASR
  - [Enhanced] Source + target ASR → ST → target ASR

- Decoder pretrained in target language
- Incorporating additional knowledge of the target language



Datasets

- Common Voice (CV), IARPA Babel (BB), LibriSpeech (LS)

- High resource (source) ASR: CV En & Fr, LS
- Low resource (target) ASR: CV, BB
- MT: OPUS indexed datasets

	Dataset	Train	Speakers
	Source ASR		
CV	Common Voice: English	477h	15.2k
CV_{Fr}	Common Voice: French	264h	1.8k
LS	Librispeech	960h	2.3k
MC	MuST-C: En-Nl	422h	2.2k
	Target ASR		
Vi	IARPA Babel 107b-v0.7	96h	0.6k
Ht	IARPA Babel 201b-v0.2b	70h	0.3K
Pt	Common Voice v4	10h	2
Zh-CN	Common Voice v4	10h	22
Nl	Common Voice v4	7h	78
Mn	Common Voice v4	3h	4

	Dataset	En/Fr Sent.	Model
Vi	OpenSubtitles	4M/3M	Base
Ht	JW300	220K/220K	Base 3+3
Pt	OpenSubtitles	33M/23M	Big
Zh	MultiUN	10M/10M	Big
Nl	OpenSubtitles	37M/25M	Big
Mn	JW300+GNOME+QED	210K/203K	Base 3+3
Nl_W	WikiMatrix	511K/-	Base 3+3
Nl_S	OpenSubtitles	37M/-	Base 3+3
Nl_M	OpenSubtitles	37M/-	Base

ST-Enhanced Cross-Lingual Transfer

- Source ASR: monolingual (En) & multilingual (En+Fr)
- Up to 24.6% WER reduction on Common Voice and IARPA Babel
- Up to 8.9% WER reduction with pseudo ST labels from low-resource MT

		Vi	Ht	Pt	Zh-CN	Nl	Mn
Baseline		57.2	66.1	62.3	90.3	96.5	109.7
From English							
CV	Src ASR	53.7	60.7	40.9	41.3	44.2	67.7
	+ ST	52.5 (-2.2%)	59.3 (-2.3%)	33.7 (-17.6%)	35.3 (-14.5%)	42.0 (-5.0%)	64.1 (-5.3%)
	Src+Tgt ASR	51.6	58.1	34.7	37.0	42.5	63.0
	+ ST	51.2 (-0.8%)	57.2 (-1.5%)	31.2 (-10.1%)	35.2 (-4.9%)	40.4 (-4.9%)	62.3 (-1.1%)
CV+LS	Src ASR	54.7	59.9	41.3	40.0	42.2	66.1
	+ ST	52.9 (-3.3%)	57.4 (-4.2%)	31.8 (-23.0%)	35.7 (-4.2%)	37.9 (-10.2%)	60.2 (-8.9%)
	Src+Tgt ASR	52.7	57.8	34.4	36.4	41.7	67.9
	+ ST	52.2 (-0.9%)	57.2 (-1.0%)	31.2 (-9.3%)	35.5 (-2.5%)	38.8 (-7.0%)	62.5 (-8.0%)
From English+French							
CV+CV$_{Fr}$	Src ASR	54.5	59.4	39.5	39.2	43.0	67.7
	+ ST	51.7 (-5.1%)	57.8 (-2.7%)	29.8 (-24.6%)	33.6 (-14.3%)	38.4 (-10.7%)	62.1 (-8.3%)
	Src+Tgt ASR	52.9	57.1	31.7	36.4	40.7	62.4
	+ ST	52.0 (-1.7%)	55.7 (-2.5%)	28.6 (-9.8%)	32.9 (-9.6%)	38.3 (-5.9%)	59.6 (-4.5%)

MT Models for Pseudo-Labeling

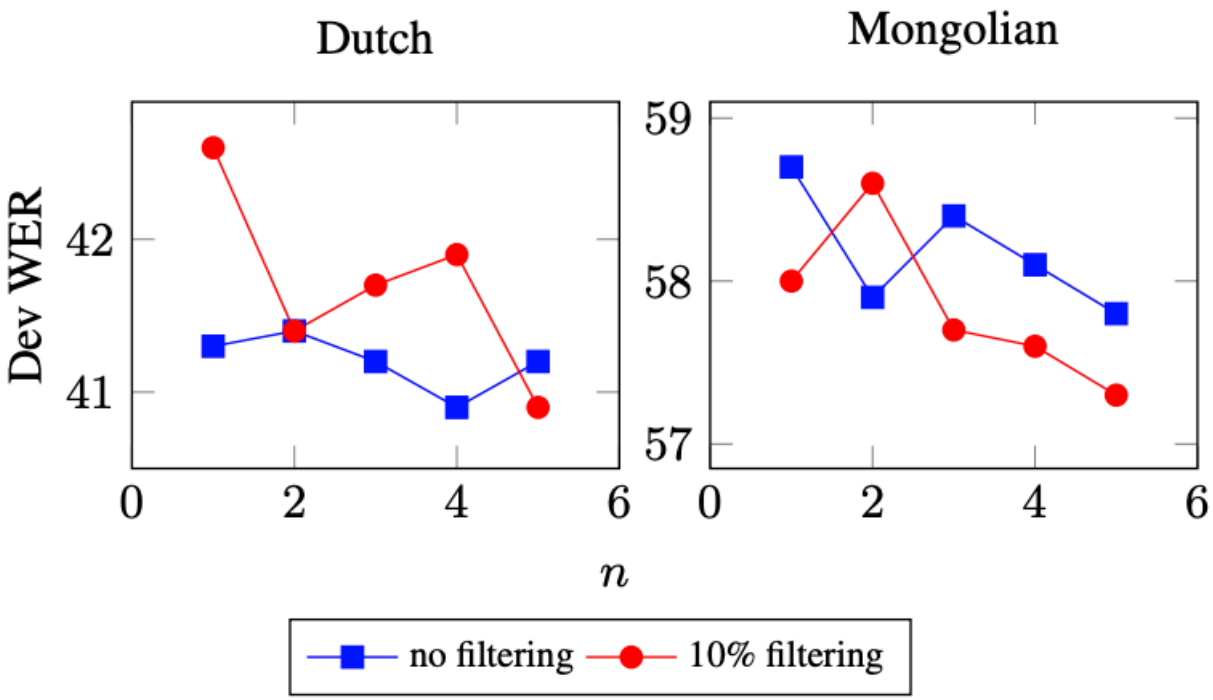
- How MT labels affect ST-enhanced transfer?
- MuST-C En-Nl ASR/ST
- Human labels from MuST-C
- MT labels from
  - Nl_W: Small model on WikiMatrix (o.5M examples)
  - Nl_S , Nl_M, and Nl: small, medium, and large model on OpenSubtitles (noisy, 37M)

	ST Label (NA for baseline transfer)					
	NA	Nl_W	Nl_S	Nl_M	Nl	Real
MT	-	24.8	34.0	34.1	35.6	**100.0**
ST	-	18.9	23.7	23.9	**24.0**	23.9
+CV	-	18.6	**23.3**	22.6	23.1	-
ASR	44.7	**42.4**	43.1	43.2	43.9	43.9
+CV	42.4	**38.7**	40.0	39.2	**38.7**	-

Pseudo-Label Sampling and Filtering

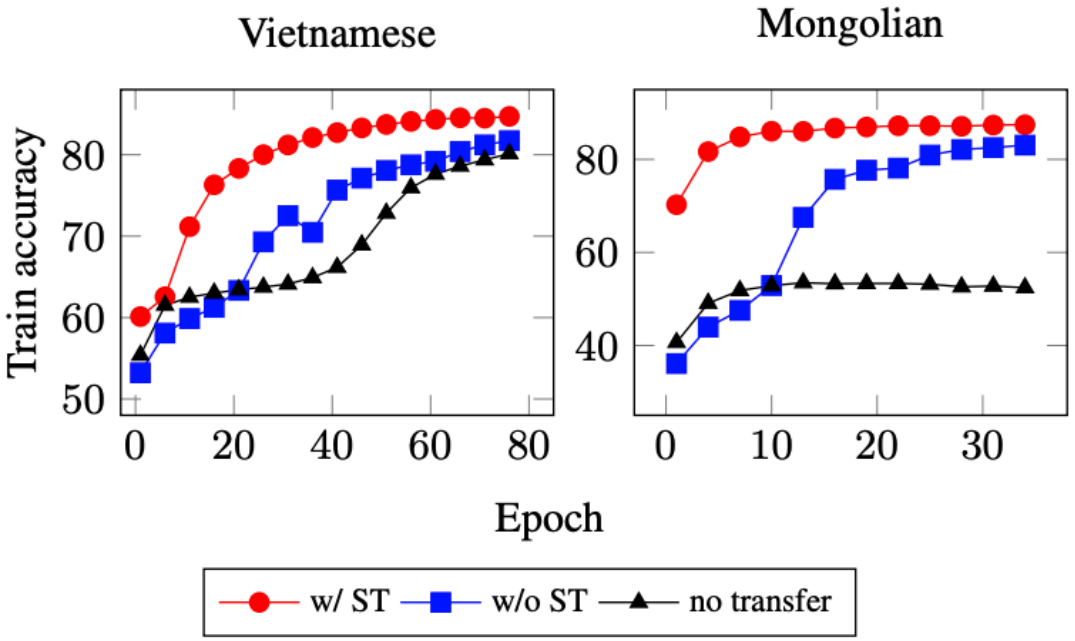
- Pseudo-labels (PL) from beam search decoding
- K predictions given beam size K
- Sample PLs in each epoch from the top N (N<=K) to alleviate overfitting

- PLs from low-resource MT may be of low quality
- Filter examples with PLs of low confidence



Effectiveness of ST Pre-training

- ST decoder trained in target language
- “Closer” to target ASR than source ASR (decoder trained in source languages)
- ST → target ASR is supposed to be faster than source ASR → target ASR



ST without ASR Pre-training

- Simplified one-step transfer: ST → target ASR
- Still brings gains in most cases
- Lack of source ASR pre-training brings difficulties to ST training

	Vi	Ht	Pt	Zh-CN	Nl	Mn
ASR	54.5	59.4	39.5	39.2	43.0	67.7
ASR→ST	51.7	57.8	29.8	33.6	38.4	62.1
ST	53.7	58.7	32.5	35.3	<u>44.1</u>	67.3

Thank you!

FACEBOOK