Speech Recognition beyond isolated English sentences

A Tale of Two Projects

Desh Raj

Center for Language and Speech Processing

draj@cs.jhu.edu

Streaming multi-talker speech recognition

[Work done during internship at Microsoft; ICASSP 2022]

Motivation

- Speech recognition systems work very well on isolated single-speaker utterances (<2% WER on LibriSpeech [3]).
- But what about conversational situations, such as meetings? Many speakers, **overlapping speech**, noise and reverberations →WER on AMI SDM >30%!
- How to design end-to-end ASR systems that can perform *streaming* transcription in such cases?

Method

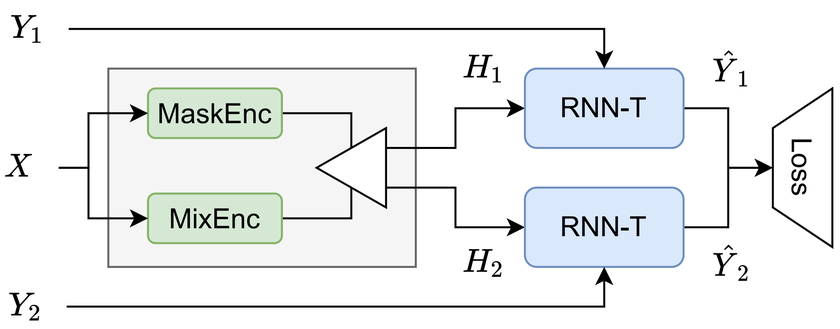

Figure 1: Streaming Unmixing and Recognition Transducer (SURT)

- The conventional approach to perform multi-talker ASR is by adding a speech separation front-end. This requires training a separate module, and also adds distortions to the speech which degrades performance.
- Instead, we extend the popular RNN-Transducer models to multi-talker scenario by adding an "unmixing" component [2], and train it end-to-end on simulated meetings.
- Since sessions can be arbitrarily long, we use dual-path models (see figure) to handle the problem of huge sequence lengths.

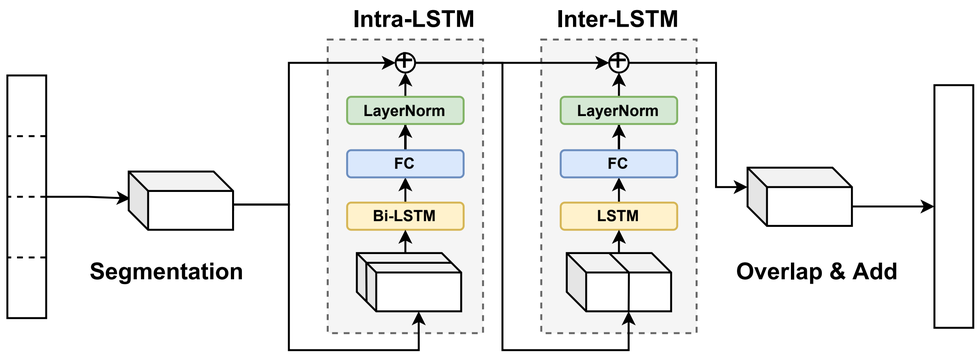

Figure 2: Dual-path modeling for long sequences

Results

- We obtain comparable results to *offline* modular systems while being *smaller* and *streaming*.

Model	# params	Overlap ratio in %				
		0	10	20	30	40
BLSTM CSS + Hybrid ASR	-	17.6	20.9	26.1	32.6	36.1
Conformer CSS + E2E ASR	197.0	6.9	9.1	12.5	16.7	19.3
SURT w/ DP-LSTM	65.4	21.6	21.7	20.6	25.4	28.4
SURT w/ DP-Transformer	42.9	18.9	19.6	21.9	23.9	28.7

Table 1: WERs on LibriCSS with increasing overlapping speech ratio in meetings

- By training the dual path models with *chunk width randomization*, we can reduce the **decoding latency from 350 ms to 150 ms** without significantly degrading WER.

Few-shot learning for low-resource languages

[Work led by Matthew Wiesner (now at COE); ICASSP 2022]

Motivation

- Self-supervised models such as Wav2vec 2.0 have recently demonstrated SOTA performance on popular Enlgish ASR benchmarks (such as LibriSpeech).
- Can we fine-tune these models on very low resource languages (such as Pashto or Haitian)?
- How can we exploit other resources (extra text, cross-lingual annotated speech) in such few-shot fine tuning?

Method

- Usually, when you want to fine tune an SSL model for end-to-end ASR, you would replace the output with a new linear layer (# nodes = # BPE units). But this only works if you have at least 30-40 hours of annotated speech in the new domain!
- Instead, we leverage *shared phonemic outputs* by using annotated speech in other languages (think International Phonetic Alphabet).
- We use extra in-domain text to create an expanded denominator graph so that we can use *sequence-discriminative training* (LF-MMI) instead of CTC.

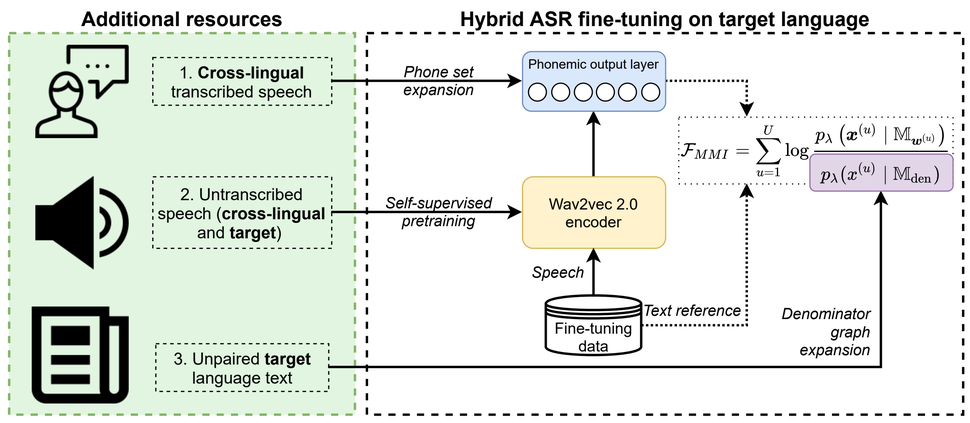

Figure 3: Injecting additional resources for ASR fine-tuning of self-supervised models

Results

- XLSR-53 + BABEL pretraining + shared phone set achieves similar WER as 80h supervised model with **just 15 minutes** of annotated data.

Model	Pre-training (h)	pus	hat	kat
Topline (80h supervised)	0 / 0	50.1	56.6	49.5
BABEL supervised pre-training	0 / 1.2k	86.2	81.5	89.2
+ shared phone set	0 / 1.2k	75.6	72.9	73.1
XLSR-53 (self-supervised) [1]	56k / 0	75.9	74.7	81.5
XLSR-53 + BABEL pre-training	56k / 1.2k	71.0	67.8	64.5
+ shared phone set	56k / 1.2k	**63.6**	**58.9**	**58.2**

Table 2: WERs on low-resource BABEL languages with **15 minutes** supervised fine-tuning

- Denominator graph expansion with extra text gives improvement across all languages.

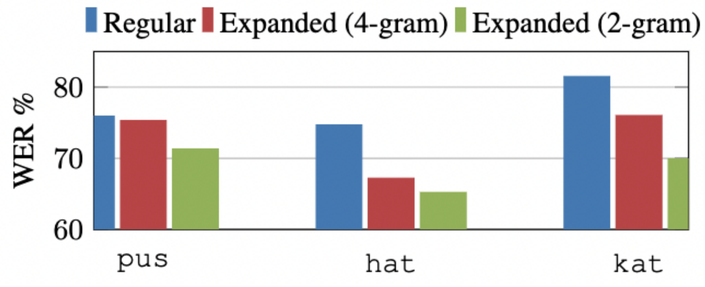

Figure 4: Effect of denominator graph expansion on LF-MMI fine-tuning

Conclusions

We can use advances in single-talker ASR with high-resource languages (transducer models, self-supervised learning) to benefit adjacent scenarios such as multi-talker conversations or low-resource languages.

References

[1] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdel rahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *ArXiv*, abs/2006.13979, 2021.

[2] Liang Lu, Naoyuki Kanda, Jinyu Li, and Yifan Gong. Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters*, 28:803–807, 2021.

[3] Yu Zhang et al. Pushing the limits of semi-supervised learning for automatic speech recognition. *ArXiv*, 2020.