Disentangling Speakers in Multi-Talker Speech Recognition with Speaker-Aware CTC

Jiawen Kang, Lingwei Meng, Mingyu Cui, Yuejiao Wang, Xixin Wu, Xunying Liu, Helen Meng
The Chinese University of Hong Kong

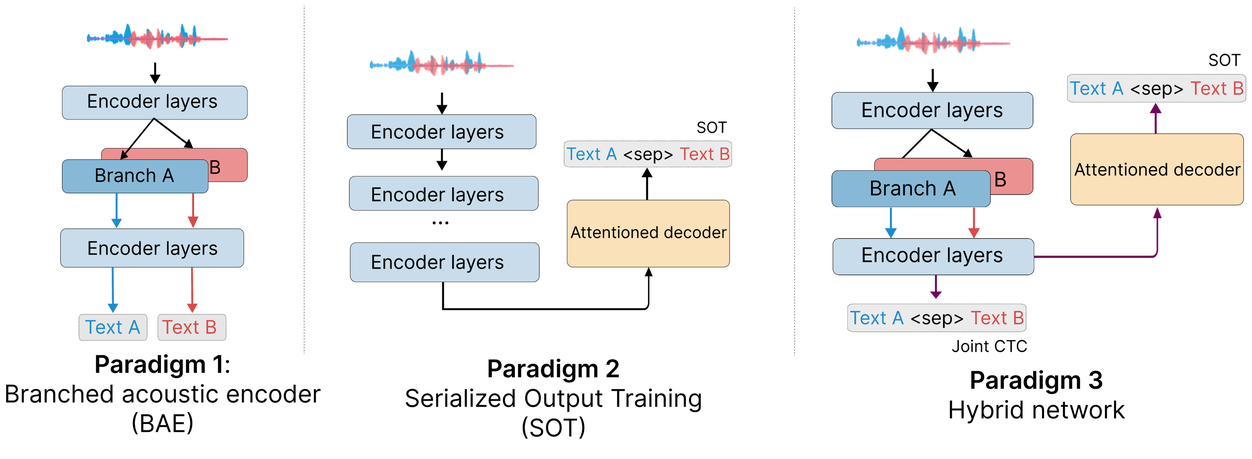
Highlights

- Insight: CTC guides encoder to represent speakers in distinct temporal regions
- A novel aspect of handling overlap disentangling: Propose a Speaker-aware CTC (SACTC) that explicitly models speaker disentanglement
- Significantly outperform baselines: relatively 10% overall WER reduction, 15% on high-overlap speech

Introduction

- Multi-talker automatic speech recognition (MTASR): Transcribe speech with multiple talkers and potential overlaps

- Existing approaches follow the following 3 paradigms:



Paradigm 1: Branched acoustic encoder (BAE)

Paradigm 2 Serialized Output Training (SOT)

Paradigm 3 Hybrid network

- **Questions**:
  - Any investigation about Connectionist Temporal Classification (CTC) in MTASR?
  - Why CTC help SOT-based MTASR in many works?
  - What about its monotonicity assumption?

Revisit Connectionist Temporal Classification (CTC)

- CTC objective - summing the posteriors of all possible alignments:

$$P(l|x) = \sum_{\pi \in B^{-1}(l)} p(\pi|x) \quad (1)$$

- p(π|x) is calculated by jointing frame-wise posteriors p(πt|xt) within a path:

$$p(\pi|x) = \prod_{t=1}^{T} p(\pi_t|x_t) = \prod_{t=1}^{T} y^t_{\pi_t} \quad (2)$$

- Use the forward-backward algorithm to calculate the posterior of all paths passing a CTC lattice node:

$$\alpha(t,v) = \sum_{\substack{\pi: B(\pi_{1:t}) = B(l'_{1:v}) \\ \pi_t = l'_v}} \prod_{t'=1}^{t} y^{t'}_{\pi_{t'}} \quad (3)$$

$$\beta(t,v) = \sum_{\substack{\pi: B(\pi_{t:T}) = B(l'_{v:2U+1}) \\ \pi_t = l'_v}} \prod_{t'=t}^{T} y^{t'}_{\pi_{t'}} \quad (4)$$

- Enumerating all possible tokens in label target sequence to obtain the posterior of all possible paths:

$$P(l|x) = \sum_{\pi \in B^{-1}(l)} p(\pi|x) = \sum_{v=1}^{2U+1} \frac{\alpha(t,v) \cdot \beta(t,v)}{y^t_{l'_v}} \quad (5)$$

- More details can be found in the paper -

Speaker-Aware CTC Based on Minimizing Bayes Risk

- Introducing Bayes risk ***r(π)***over the posteriors of each alignment path:

$$\mathcal{J}_{br}(l,x) = \sum_{\pi \in B^{-1}(l)} r(\pi) \cdot p(\pi|x) \quad (6)$$

- Design a speaker-aware risk function to constrain the emitting times of tokens with consideration of their belonging speakers:

$$r_{sa}(s,t) = \begin{cases} -\frac{1}{1+e^{(\lambda(\frac{t}{T}-b))}} & \text{if } s = 1 \\ -\frac{1}{1+e^{-(\lambda(\frac{t}{T}-b))}} & \text{Otherwise} \end{cases}, \quad b = \frac{M}{M+N} \quad (7)$$

- where ***s*** represents the speaker index ordered by first-in-first-out, ***b*** is a ratio of utterance lengths, used to determine a speaker boundary
- The ***rsa(s, t)*** overall is a conditional Sigmoid function, assigning high or low risks according to the certain speaker boundary

- For a certain token ***u*** of a speaker ***s***, summarizing Bayes risks of all time frames:

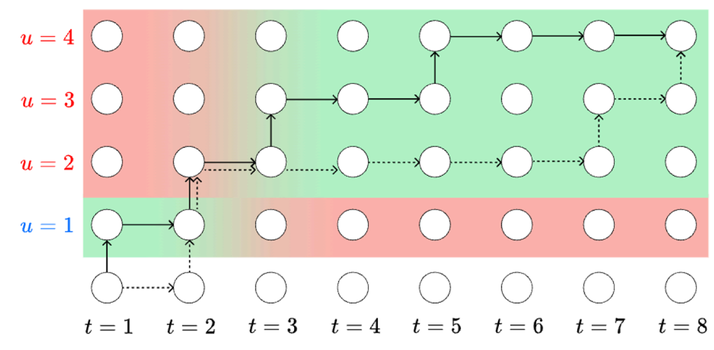
$$\mathcal{J}'_{sa}(l,x,s,u) = \sum_{t=1}^{T} r_{sa}(s,t) \cdot \frac{\alpha(t,2u) \cdot \hat{\beta}(t,2u)}{y^t_{\pi_t}} \quad (8)$$

- Enumerating all tokens from all speakers, the final training objective is to minimize the following Bayes risk:

$$\mathcal{J}_{sa}(l,x) = \frac{1}{S} \cdot \sum_{s=1}^{S} [\frac{1}{U} \cdot \sum_{u}^{U} \mathbb{1}(s,u) \cdot log\mathcal{J}'_{sa}(l,x,s,u)] \quad (9)$$

- Detailed derivation can be found in the paper -

- A simplified illustration of the proposed speaker-aware risk function with CTC lattice:

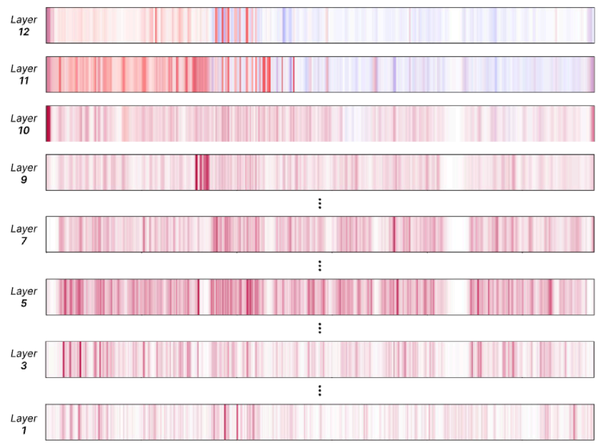


- Red area: high risk
- Green area: low risk
- Tokens 1 and 2,3,4 are from two different speakers
- Two encouraged alignments are shown as examples.

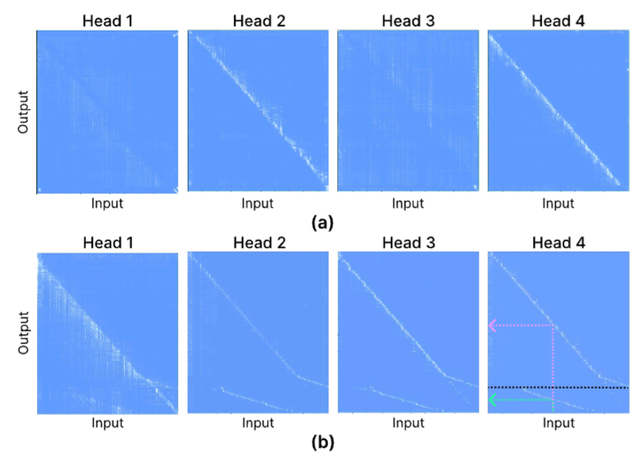
Experiments

- **Settings**:
  - Dataset: LibriSpeechMix (LSM) with 1,700 hours of speech
  - ASR model: Conformer-based attention-encoder-decoder (AED) architecture
  - Metrics: word error rate (WER) & overlap-aware WER (OA-WER)
  - More details can be found in the paper

- **Analysis of CTC**:
  - Left: The encoder learns to disentangle speakers temporally
  - Right: CTC reorders speaker information to align with concatenated labels.



top-50 attended frames in SOT-CTC ASR encoder for two speakers (red and blue colors)



Attention matrices in the last conformer blocks of (a) SOT and (b) SOT-CTC models.

- **Main results**

TABLE I
WER (%) OF VANILLA CTC AND SACTC IN MTASR. C1 AND E1 ARE THE MAIN EXPERIMENTS. "DEC." STANDS FOR DECODING.

ID	System	Librispeech		LibrispeechMix-2mix					
		dev	test	Dev	Test (Overall)	Test (Conditional)			
						low	mid	high	OA-WER
A1	SOT	4.1	4.6	7.9	9.2	9.0	**8.0**	12.8	9.9
B1	CTC	5.0	5.4	11.7	11.1	7.5	12.4	18.2	12.7
C1	SOT+CTC	4.3	4.5	8.4	8.8	7.1	9.0	13.1	9.7
C2	↪ AED only dec.	4.8	5.4	11.5	12.9	11.7	12.5	17.3	13.8
C3	↪ CTC only dec.	5.5	5.6	12.7	12.0	7.8	13.8	19.7	13.8
D1	SACTC	5.4	5.6	13.5	12.3	8.0	13.9	20.8	14.2
E1	SOT+SACTC	3.9	**4.1**	8.2	**8.0**	**6.0**	8.4	12.8	**9.1**
E2	↪ AED only dec.	4.0	4.5	8.4	8.8	8.2	**8.0**	**12.3**	9.5
E3	↪ CTC only dec.	5.5	5.8	12.5	11.9	8.1	13.1	19.9	13.7

- CTC is better on low-overlap speech, worse overall
- SOT+CTC consistently outperform SOT
- SACTC alone doesn't outperform vanilla CTC
  - SACTC doesn't directly optimize recognition results, but regularize embeddings
- SOT+SACTC consistently outperform SOT+CTC
- Decoding with AED only leads to the best high-overlap performance

- **Other experiments**
  - Adopt different risk factors (5~20) yield consistent improvement over baselines
  - SOT-SACTC could generalize to 3 speaker seniors, even though there are only 1-speaker and 2-speakers speech in training data

Conclusions

- CTC could be used for speaker disentanglement
- A tailor-made speaker-aware CTC (SACTC) was further proposed to explicitly do this
- The SOT-SACTC significant (10%~20% relative improvement) outperforms all baselines
- This might be a new perspective on speaker separation

- Future works:
  - Extending to streaming scenarios
  - Exploring non-autoregressive multi-talker speech recognition

GitHub arXiv

Contact: jwkang@se.cuhk.edu.hk