Augmenting Transformer-Transducer Based Speaker Change Detection With Token-Level Training Loss

Guanlong Zhao (guanlongzhao@google.com), Quan Wang, Han Lu, Yiling Huang, Ignacio Lopez Moreno

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

Introduction

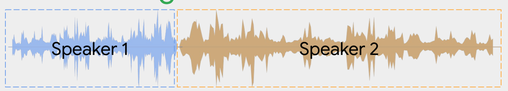
Problem statement: Perform word-level Speaker Change Detection (SCD) with a Transformer-Transducer model

Challenges

- Speaker turns are sparse compared to regular spoken words -- one speaker turn per 40+ words
- Suboptimal evaluation metrics

Solutions

- Token-based training loss + interval-based eval metrics

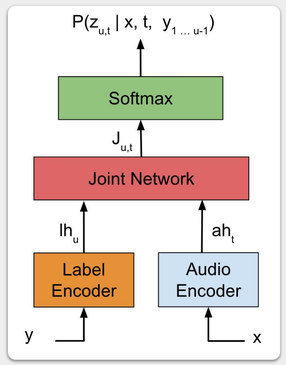


What's the weather? <st> It's sunny outside.

System Description

Baseline SCD model

- We treat speaker turn as a new special token <st>
- Jointly trained with the ASR model
- Audio encoder: 15 layers of transformer blocks
- Output: 75 possible graphemes (including <st>, <sos>, <eos>)



Token-based training loss

- The idea is to minimize the expected FA and FR rates of the <st> token in the prediction
- Achieved by a customized minimum edit distance alignment

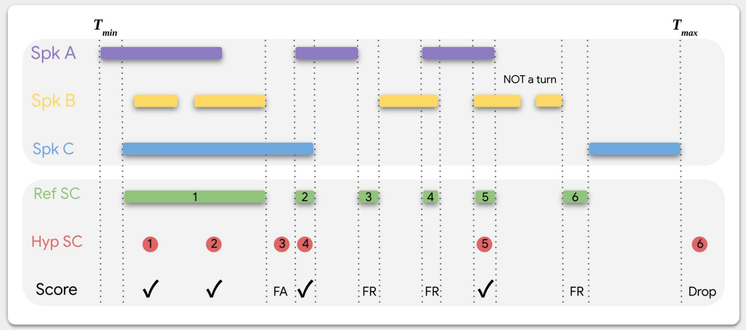
$$\text{sub-cost}(r,h) = \begin{cases} 0, & \text{If } r = h; \\ 1, & \text{If } r \neq h \neq <\text{st}>; \\ +\infty, & \text{Otherwise.} \end{cases}$$

$$\text{ins/del-cost}(token) = \begin{cases} k \geq 1, & \text{If } token = <\text{st}>; \\ 1, & \text{Otherwise.} \end{cases}$$

- Conceptually, the training loss is

$$L_{\text{SCD}} = \alpha \cdot \text{WER} + \beta \cdot \text{FA}_{\text{SCD}} + \gamma \cdot \text{FR}_{\text{SCD}} - \lambda \log P(Y|X)$$

Evaluation Metrics



Conventional evaluation metrics

- Timestamp-based precision and recall rates: sensitive to inaccurate annotations and deviations of timestamps
- Purity and coverage: indirect SCD quality measurements

Proposed interval-based precision and recall: proper handling of overlapping speech

- Assumption: dense speaker label annotations
- Treat speaker changes as intervals rather than points
- Find the time intervals that speaker changes happen, e.g., overlapping speech segments imply speaker turns
- Find SCD predictions that fall into these intervals
- Compute the precision and recall rates accordingly

Experimental Setup

Datasets

- Train: Fisher, Callhome English, AMI, ICSI, internal long-form sets

Testset	Domain	Dur. (h)	Average	
			Turns/min	Dur./Rec. (min)
AMI	Meeting	9.1	10	34
Callhome	Telephone	1.7	19	5
DIHARD1	Mixed	16.2	12	9
Fisher	Telephone	28.7	13	10
ICSI	Meeting	2.8	13	55
Inbound	Telephone	21.0	9	5
Outbound	Telephone	45.6	13	6

Systems

- **Baseline**: Trained with the negative log probability loss
- **EMBR**: Baseline + EMBR loss
- **SCD loss (proposed)**: Baseline + proposed training loss
- All share the same architecture (27M parameters)

Results

Long-form results

- F1 of proposed precision and recall
  - **SCD loss** vs. **Baseline**: +8.9% relative
    - +16.8% relative recall
    - Comparative precision (-0.6% relative)
  - **SCD loss** vs. **EMBR**: +3.5% relative
- F1 of timestamp-based precision and recall rates: low absolute values; +13.4% relative compared with **Baseline**
- F1 of purity and coverage: all comparable

Evaluation Metric	System	AMI	CallHome	DIHARD1	Fisher	ICSI	Inbound	Outbound	Pooled data
Precision (%)	Baseline	80.9	81.0	78.7	81.8	78.7	73.0	76.3	78.1
	EMBR	**81.3**	**82.0**	**79.8**	**83.5**	**79.3**	**74.3**	**77.0**	**79.1**
	SCD loss	79.4	**82.0**	78.8	82.6	77.8	72.8	75.1	77.6
Recall (%)	Baseline	64.0	50.6	49.2	62.4	54.3	62.2	50.9	55.8
	EMBR	64.2	53.4	49.5	71.1	53.6	71.8	53.6	60.3
	SCD loss	**68.1**	**59.1**	**52.4**	**75.7**	**58.7**	**79.2**	**58.7**	**65.2**
F1 (%) (Precision & Recall)	Baseline	71.5	62.3	60.6	70.8	64.2	67.2	61.1	65.1
	EMBR	71.7	64.7	61.1	76.8	64.0	73.0	63.2	68.5
	SCD loss	**73.3**	**68.7**	**62.9**	**79.0**	**66.9**	**75.9**	**65.9**	**70.9**
Purity (%)	Baseline	87.4	84.3	90.3	80.5	76.9	95.0	76.7	82.7
	EMBR	87.6	84.1	90.5	82.7	77.0	95.3	77.1	83.5
	SCD loss	**88.5**	**84.9**	**91.0**	**83.5**	**77.7**	**95.5**	**78.3**	**84.3**
Coverage (%)	Baseline	**70.0**	**85.6**	64.9	**80.8**	79.3	**77.1**	83.4	**78.5**
	EMBR	**70.0**	85.3	**65.1**	80.6	**79.8**	76.7	**83.7**	**78.5**
	SCD loss	68.7	84.7	64.7	80.2	78.9	75.0	82.4	77.5
F1 (%) (Purity & Coverage)	Baseline	**77.8**	**84.9**	75.6	80.6	78.1	**85.1**	79.9	80.5
	EMBR	**77.8**	84.7	**75.7**	81.6	**78.4**	85.0	**80.3**	**80.9**
	SCD loss	77.3	84.8	75.6	**81.9**	78.3	84.0	**80.3**	80.8

Short-form results

- Segmented from the long-form data
- Focuses on short utterances quality
- Similar trend as in long-form

Length	System	F1 (Precision & Recall)	F1 (Purity & Coverage)
30s	Baseline	55.2	75.9
	EMBR	60.9	80.8
	SCD loss	**65.0**	**81.5**
60s	Baseline	58.6	77.9
	EMBR	64.4	**81.1**
	SCD loss	**67.9**	**81.1**
120s	Baseline	61.8	79.5
	EMBR	66.6	**81.2**
	SCD loss	**69.6**	81.0

Additional Resources

Supplemental results

Google AI Blog post

Recorder App on Pixel