USM-SCD: Multilingual Speaker Change Detection Based on Large Pretrained Foundation Models

Guanlong Zhao (guanlongzhao@google.com), Yongqiang Wang, Jason Pelecanos, Yu Zhang, Hank Liao, Yiling Huang, Han Lu, Quan Wang

ICASSP 2024 KOREA

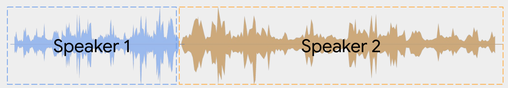
Introduction

Problem statement

Leveraging large pretrained foundation models to build a high-quality multilingual model for ASR and SCD

Contributions

- A 96-language SCD model that significantly outperforms previous baselines
- **75.3%** average SCD F1 score across 96 languages
- **85.8%** SCD F1 score on En-US



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

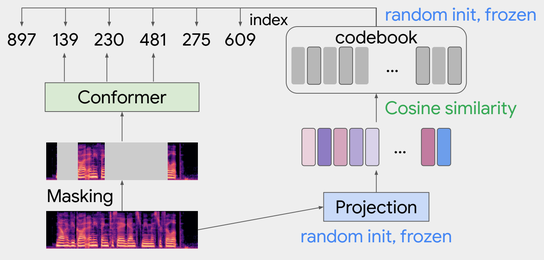
Method

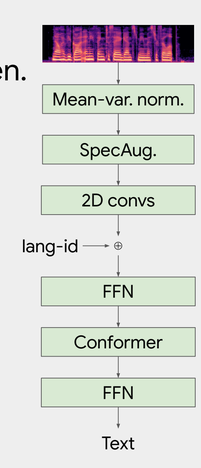
Backbone model

- Encoder: Conformer with chunk-wise atten.
- Decoder: CTC on WordPiece tokens

Pretraining

- Unsupervised BEST-RQ pretraining





- Supervised ASR pretraining: Init from BEST-RQ and fine-tune on the ASR data to predict text from audio

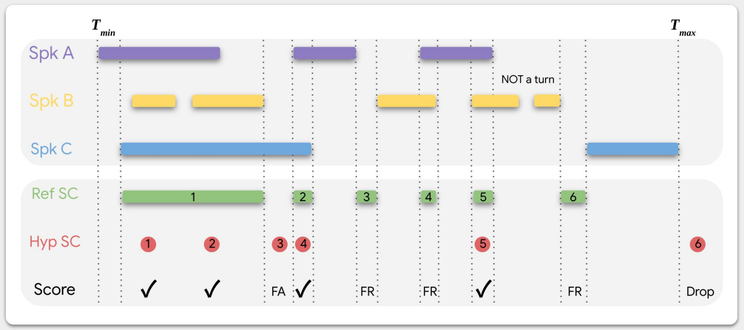
USM-SCD fine-tuning

- Fine-tune the pretrained model with SCD data
- Warm start the backbone model's Conformer encoder from a pretrained model's encoder
- Training data generation: Insert an SCD token <st> between the transcripts of different speakers

Speaker change token posterior scaling

$p'(\texttt{<st>}|\mathbf{X}) = \lambda \cdot p(\texttt{<st>}|\mathbf{X}), \lambda > 1$ ⬅️ Mitigates <st> sparsity

Evaluation Metrics



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

Training data

- 3M hrs YT 56-lang *unsupervised* data for BEST-RQ
- 1.3M hrs shortform VS 85-lang *supervised* data for ASR
- 108k hrs YT 96-lang *supervised* data for ASR + SCD

Eval data

- YT-96-Eval: 1.4k hrs YT 96-lang eval set, 15.2 hrs per lang
- Additional public and internal En-US eval sets: AMI, CALLHOME, DIHARD, Fisher, ICSI, internal telephony sets

Modeling details

- Frontend: 128-dim log mel-spec, 32ms frame, 10ms hop
- Output vocab: 16,384 WordPiece tokens
- # params: 1.84B (32 conformer layers)
- 30s max audio length in training

Baselines

- ASR: OpenAI Whisper large-v2, 1.55B params
- SCD: *SCD loss* system; "Augmenting transformer-transducer based speaker change detection with token-level training loss," ICASSP 2023

Results

Overall system comparisons

		BEST-RQ Pretrain w/ SCD	ASR Pretrain w/ SCD	ASR Pretrain w/o SCD	Whisper large-v2
WER	En-US	17.1	**12.6**	**12.6**	16.2
	21-lang.	21.1	**16.6**	**16.6**	30.1
	96-lang.	34.3	30.1	**28.8**	-
SCD	Precision	80.0	**82.4**	-	-
	Recall	**52.6**	51.9	-	-
	F1	63.5	**63.7**	-	-

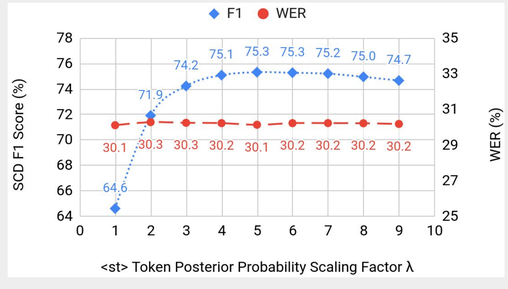
- Supervised ASR pretraining is better
- SCD leads to a **4.5%** relative WER regression (96-lang)

Effect of sub-components to fine-tune

Fine-tuned Enc. layers	# Params Trained	WER	Precision	Recall	F1
First 4	254M	35.9	83.8	35.6	50.0
Last 4	254M	30.4	82.2	44.6	57.8
First 4 & last 4	480M	**30.1**	**84.0**	**52.5**	**64.6**
All	1.84B	**30.1**	82.4	51.9	63.7

- Lower layers are more important
- Only need to tune **26%** params to maintain the quality

Effect of the speaker change token posterior scaling



- Nominal impact on WER (no scaling when λ=1)
- Best config (λ=5) SCD F1: **64.6%** -> **75.3%** (**16.6%** rel. ⬆️)

En-US deep-dive

Metrics	System	AMI	CallHome	DIHARD1	Fisher	ICSI	Inbound	Outbound	*Pooled data*
WER	SCD loss	39.8	33.0	-	30.6	46.1	-	-	33.5
	USM SCD	25.7	18.6	-	18.4	31.5	-	-	**20.7**
Precision	SCD loss	79.4	82.0	78.8	82.6	77.8	72.8	75.1	77.6
	USM SCD	91.6	84.6	92.9	94.7	90.2	94.4	91.9	**90.8**
Recall	SCD loss	68.1	59.1	52.4	75.7	58.7	79.2	58.7	65.2
	USM SCD	75.3	90.8	81.7	76.5	82.7	70.1	87.3	**81.4**
F1	SCD loss	73.3	68.7	62.9	79.0	66.9	75.9	65.9	70.9
	USM SCD	82.6	87.6	86.9	84.6	86.3	80.5	89.5	**85.8**

- Baseline is monolingual and trained on more En-US data
- Rel. SCD ⬆️: precision **17%**, recall **24.8%**, F1 **21%**