Google

On the Success and Limitations of Auxiliary Network Based Word-Level End-to-End Neural Speaker Diarization


Yiling Huang


Weiran Wang


Guanlong Zhao


Hank Liao


Wei Xia


Quan Wang

INTERSPEECH 2024

Key messages

Problem

- Can we add speaker diarization capability to **any** off-the-shelf transducer-based ASR model, with **zero WER regression**?
- Can we totally get rid of speaker **embeddings** (often considered as biometrics) and **clustering**?

Proposal

- Word-level end-to-end neural diarization (WEEND)
- Auxiliary encoder and joint network on frozen ASR model, **sharing blank logits**

Results

- Outperforms turn-to-diarize baseline on **2-speaker** and **shortform** cases
- Limitations of initial results on **multi-speaker** and **longform** cases
- Next steps: balanced **data**, increased sequence **length** in training, **DiarizationLM**

Agenda

01 Motivation

02 Related work

03 Method

04 Experiment setup

05 Results & observations

06 Conclusion & future work

Section 1

Motivation

Deploying a speaker diarization system

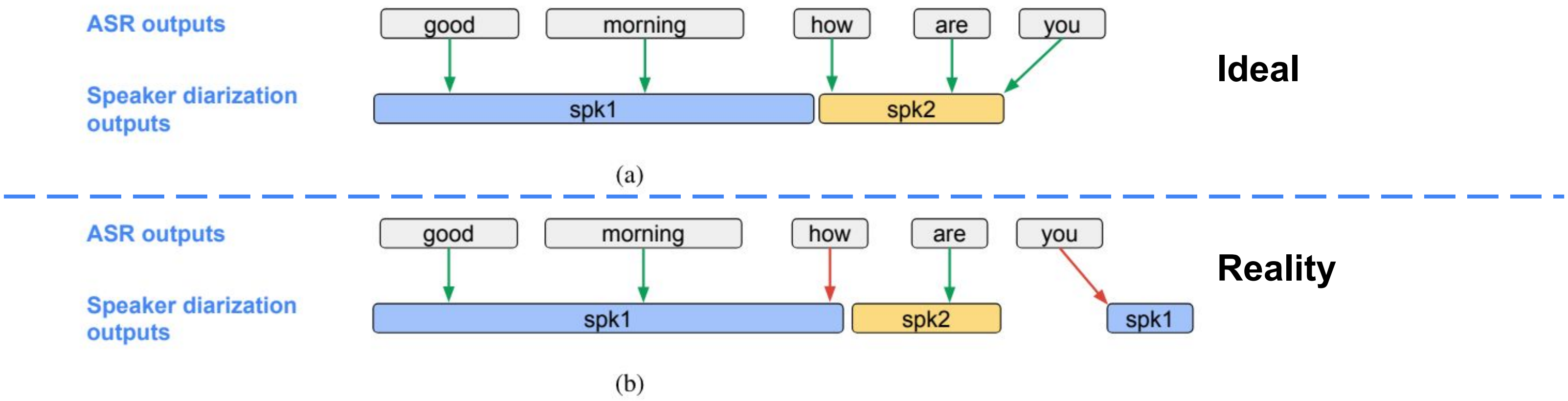
- Assume you have a speaker diarization system
- May be implemented using any algorithm:
    - D-vector / X-vector + clustering
    - End-to-End Neural Diarization (EEND) + permutation-invariant loss
    - Target-Speaker Voice Activity Detection (TS-VAD)
- You got great Diarization Error Rate (DER) on your eval sets
- Now it's time to deploy it

Challenge: ASR needed

- Speaker diarization answers the question: "who spoke **when**"
- But most realistic applications want: "who spoke **what**"
- From the "**when**" to the "**what**":
  - We need to combine speaker diarization results with ASR results
  - This is usually done by comparing timestamps

Challenge: inaccurate timestamps

- Comparing timestamps between ASR and diarization?
- This is very error-prone:
  - Two systems trained with different datasets and algorithms
  - Segmentation boundaries for diarization can be very inaccurate
  - ASR word timing inferred from the probability lattice of the decoder can be inaccurate



Common solution: joint ASR + diarization

- To solve the right problem "who spoke **what**", a common solution is to train ASR and speaker diarization together
- No need for timestamps

Section 2

Related work

RNN-Transducer and Transformer-Transducer

- RNN-T and T-T have become the mainstream end-to-end framework for speech recognition
- Allows us to introduce new **task-specific tokens** to the output vocabulary

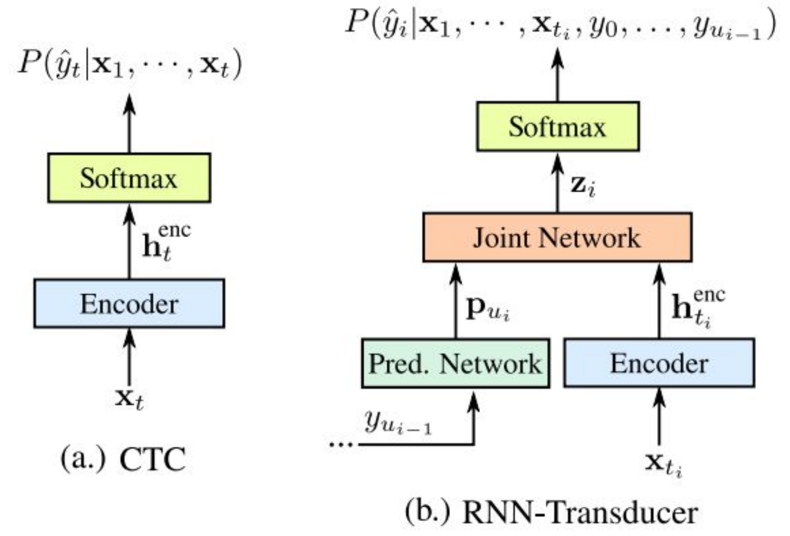


Fig. 1: A schematic representation of CTC and RNNT.

Y. He *et al.*, “Streaming end-to-end speech recognition for mobile devices”, ICASSP 2019.

Medical speaker diarization

- Introducing two new role tokens to the ASR output vocabulary:
  - <spk:dr> for doctor
  - <spk:pt> for patient
- Example output:

```
hello dr jekyll <spk:pt> hello mr hyde what brings you here today <spk:dr> I am struggling again with my bipolar disorder <spk:pt>
```

- Limitation: Role classification is **NOT** real speaker diarization

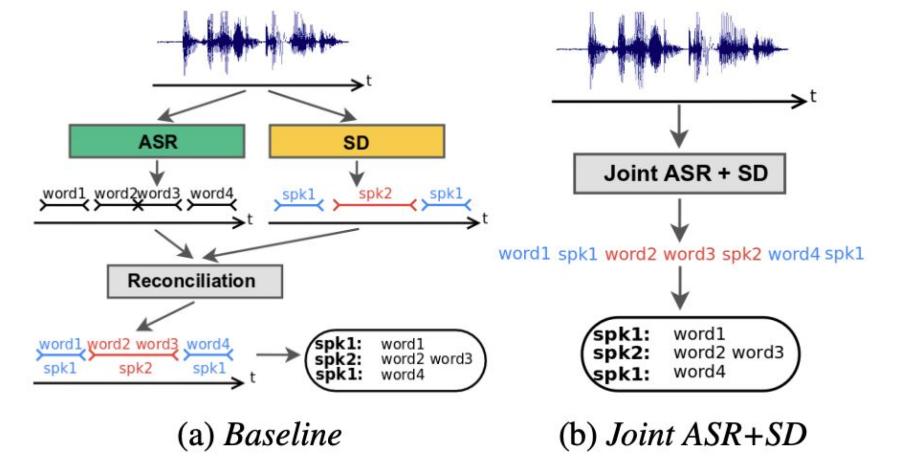
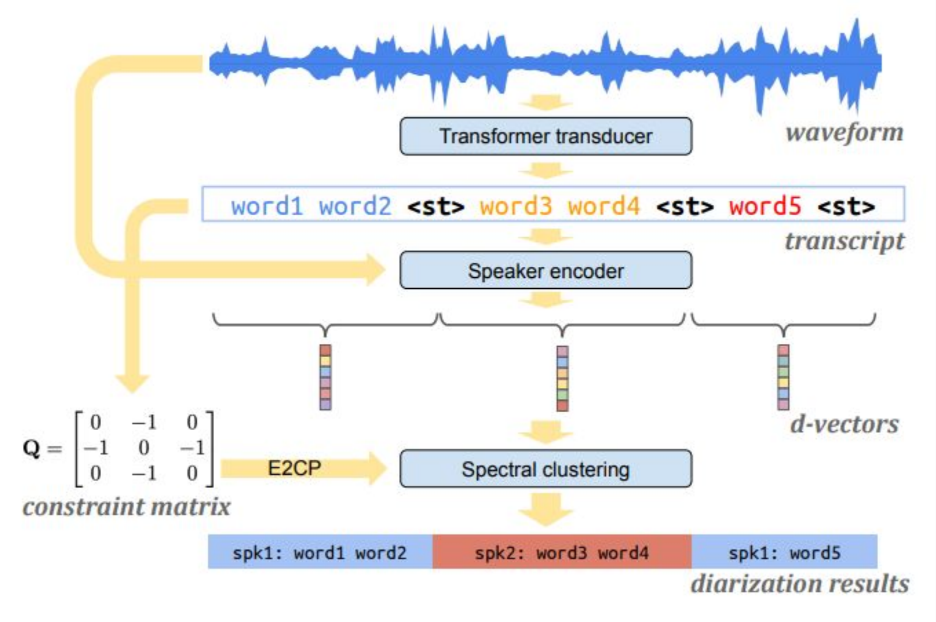


Figure 1: *Comparison of the conventional speech recognition and speaker diarization system (Figure 1a) with the proposed approach (Figure 1b), where the task consists of generating a speaker-decorated transcript from raw audio.*

L. Shafey *et al.*, "Joint Speech Recognition and Speaker Diarization via Sequence Transduction", Interspeech 2019.

Turn-to-diarize

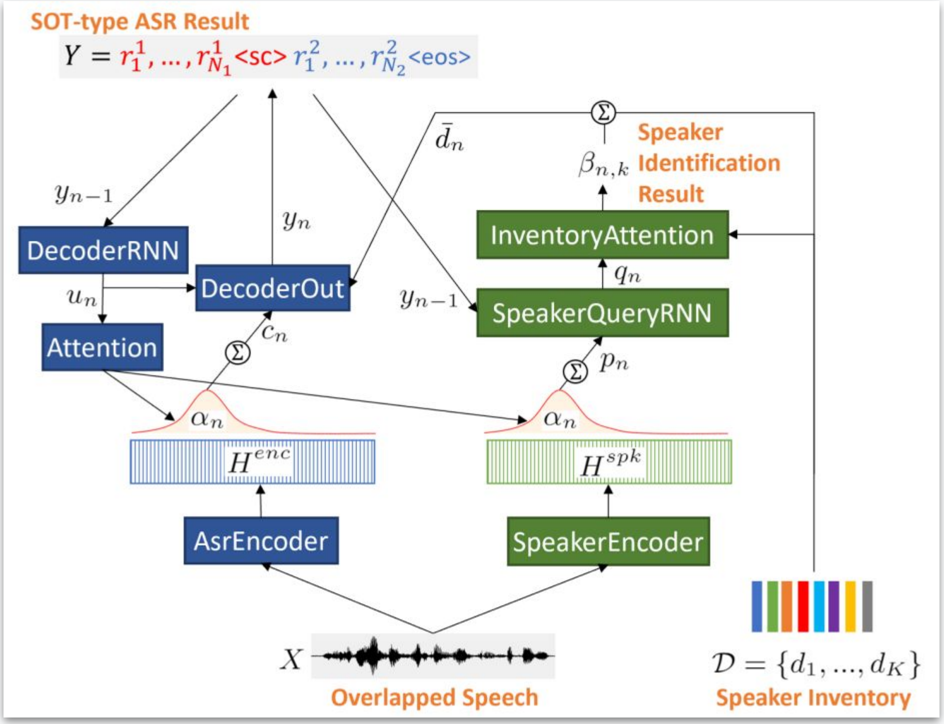
- Introducing a new speaker turn token <st> to the ASR output vocabulary
- Extract speaker embedding from each turn
- Cluster turn-level speaker embeddings



Xia, Wei, et al. "Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

Speaker Attributed ASR (SA-ASR)

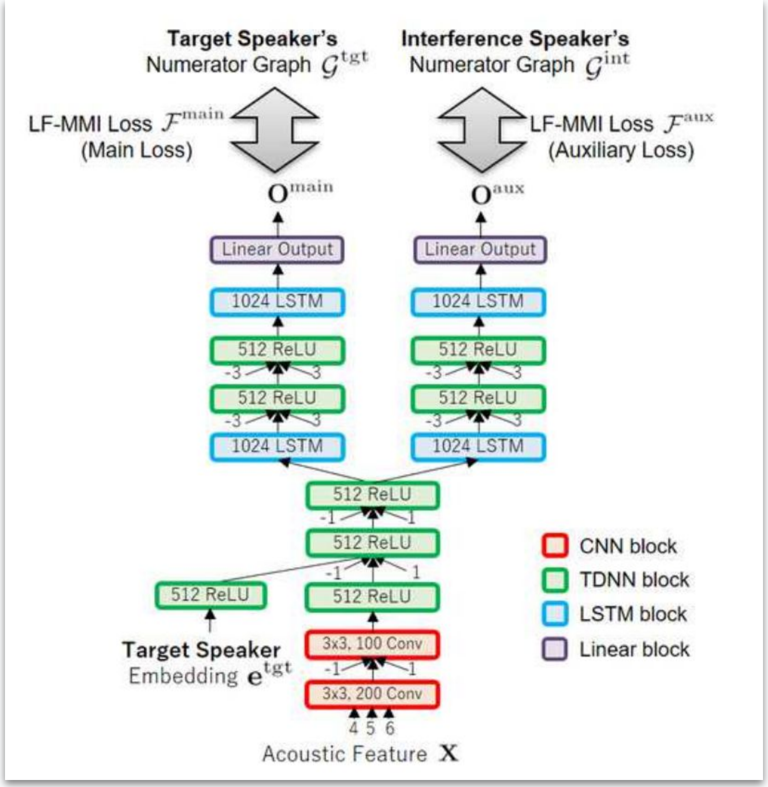
- Takes the additional inventory of speaker profiles as input
- Identifies speaker profile indices based on an attention mechanism



Kanda, Naoyuki, et al. "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers." arXiv preprint arXiv:2006.10930 (2020).

Target Speaker ASR (TS-ASR)

- Diarizing target speaker speech via enrolled speaker embedding extraction



Kanda, Naoyuki, et al. "Auxiliary interference speaker loss for target-speaker speech recognition." arXiv preprint arXiv:1906.10876 (2019).

Challenge: ASR regression

Assume we have a joint ASR and speaker diarization model, such as SA-ASR or TS-ASR

Ideal:

- The model has great DER on diarization eval sets, let's deploy it

Reality:

- We need to compare with the **best ASR baseline** model
- The Word Error Rates (WER) has a X% **regression** on some of the ASR evaluation datasets, so no

Why the gap?

- **Data:**
  - There are way more training data for ASR than for diarization
- **Model:**
  - Algorithms that works best for speaker diarization may not be best for ASR
- **People:**
  - In big companies, ASR and speaker diarization are usually owned by different teams
- **Product:**
  - In most products, ASR is a more critical feature than speaker diarization

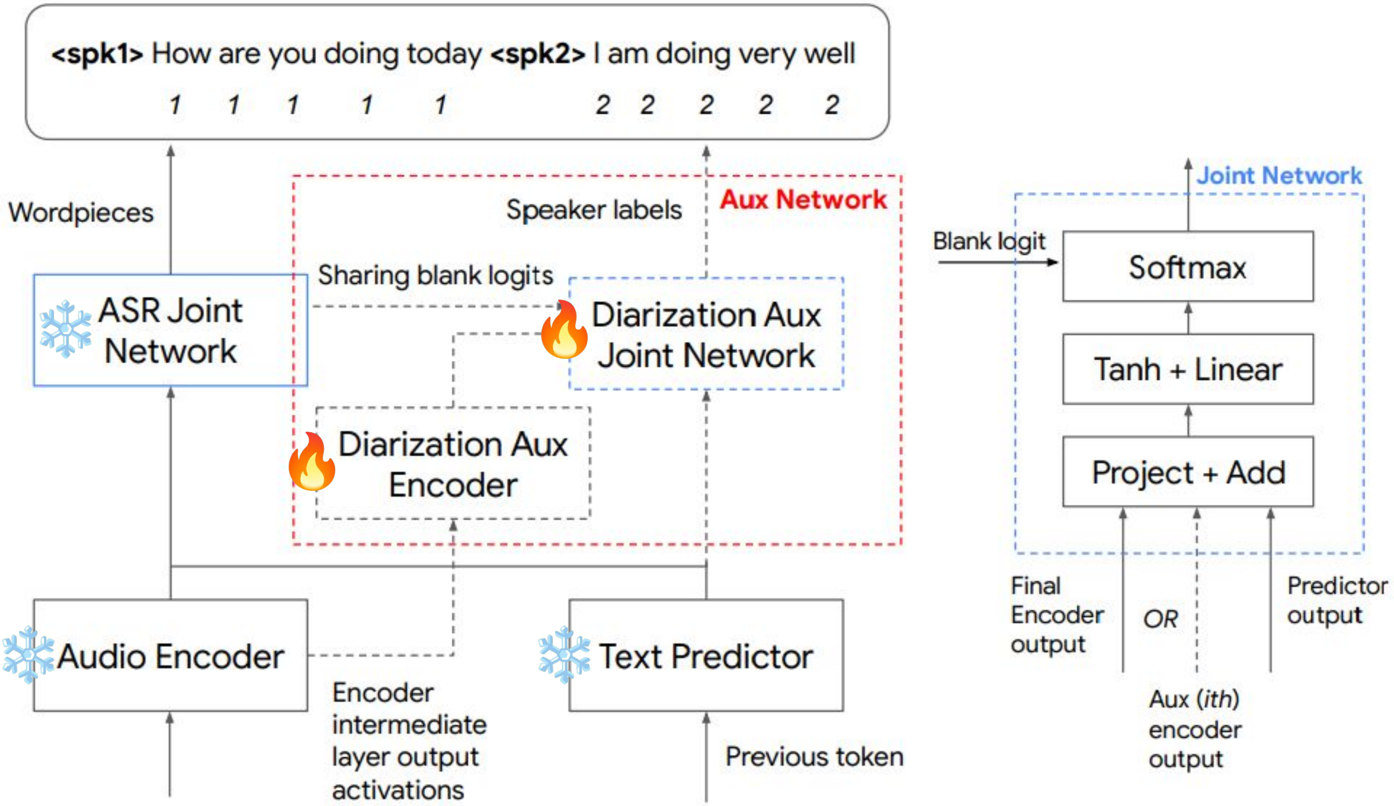
Proposal

Can we take any off-the-shelf transducer based ASR model,
add speaker diarization capability to it,
without any WER regression?

Section 3

Method

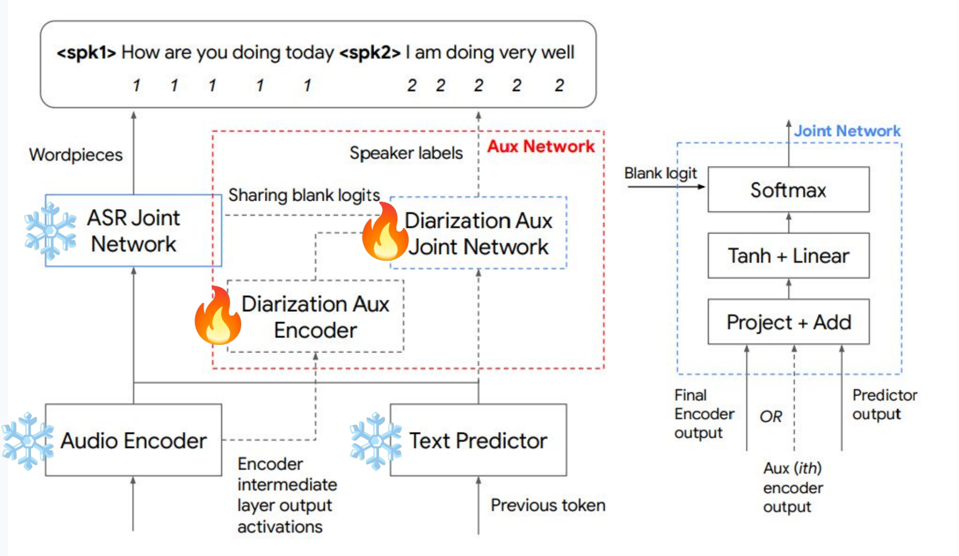
WEEND: Word-level End-to-End Neural Diarization



WEEND

Given a production-ready ASR model:

- We freeze the encoder, decoder, and joint network
- Multi-output RNN-T
- We introduce:
  - Diarization Aux Encoder
  - Diarization Aux Joint Network



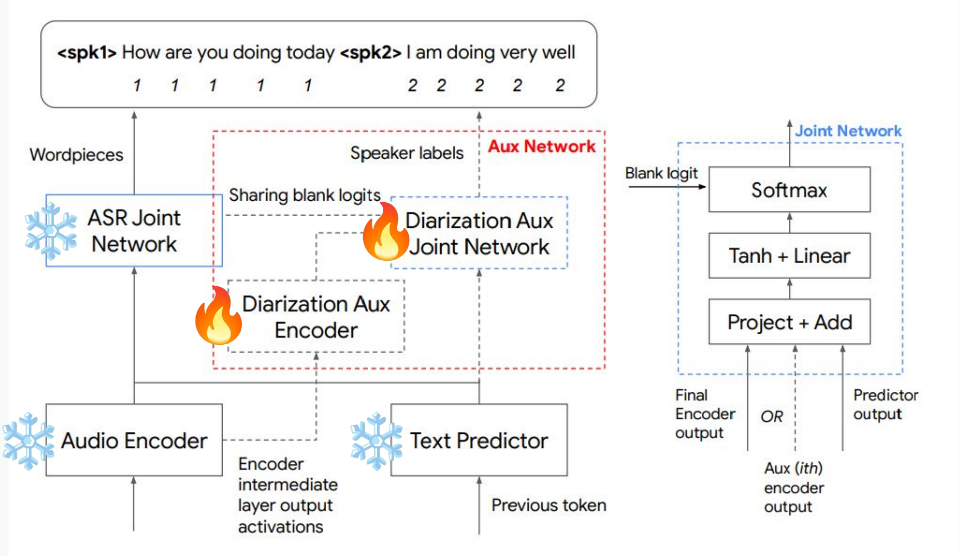
WEEND model architecture

Wang, Weiran, et al. "Multi-output RNN-T joint networks for multi-task learning of ASR and auxiliary tasks." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

WEEND

Diarization Aux Joint Network:

- Outputs sequence of integer speaker labels
- Share the **blank logits** with ASR joint network:
    - Such that each ASR word has exactly one speaker label



WEEND model architecture

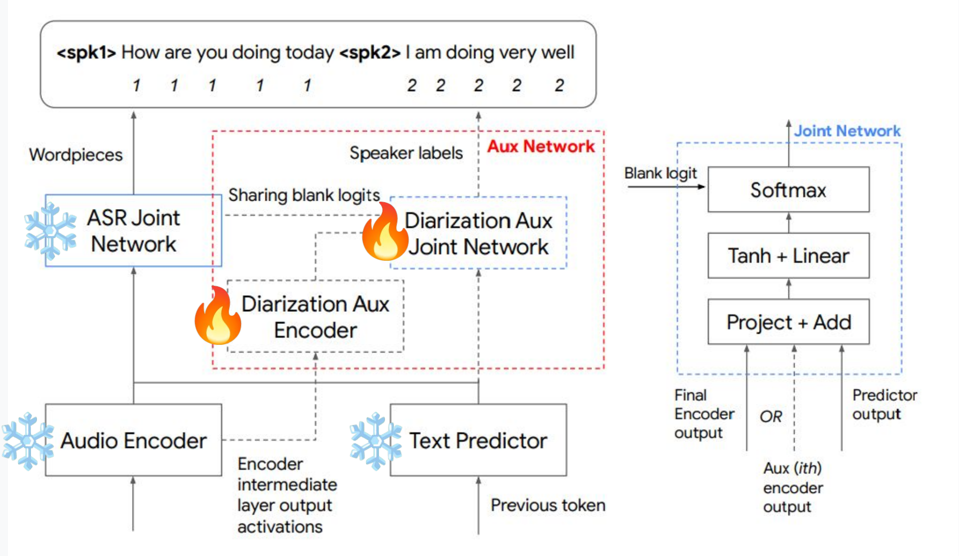
WEEND

Training on diarization data:

- We optimize RNN-T loss on the sequence of speaker labels

Deploy to production:

- Zero ASR quality regression, because ASR components are frozen



WEEND model architecture

Section 4

Experiment setup

Datasets

Public data (all English):

- AMI MixHeadset
- Callhome American English
- Fisher English

Simulated dataset:

- We concatenate utterances from LibriSpeech
- Inserted pause (0.2s ~ 1.5s)
- Applied cross-fade (0s ~ 0.2s)

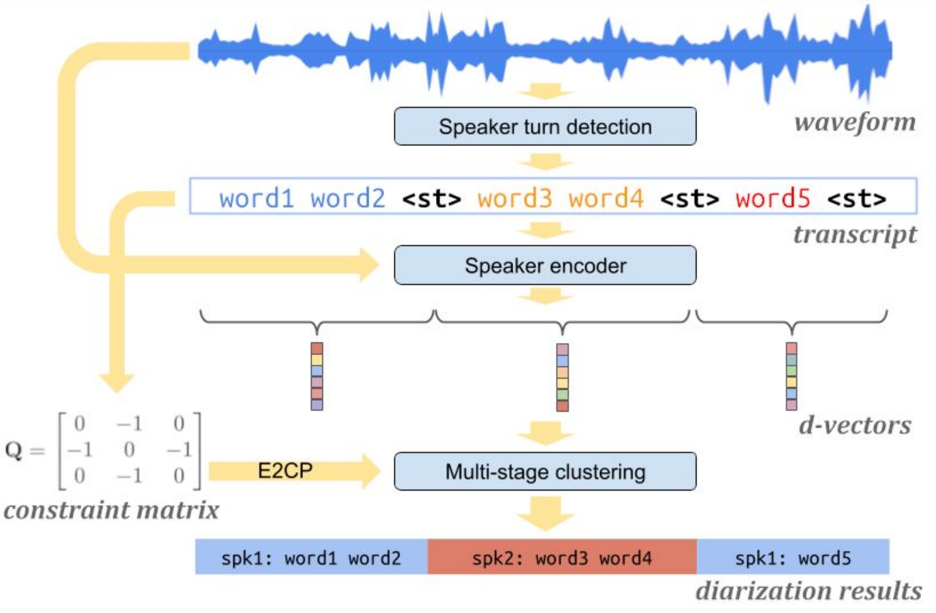
Table 1: *Diarization public and simulated datasets statistics*

Datasets	Domain	# Spk	Avg length (sec) Train	Avg length (sec) Eval	Total hours (hr) Train	Total hours (hr) Eval
AMI	Meeting	3-4	15/30/60	2039	81	9.1
Callhome	Telephone	2	15/30/60	301	14	1.7
Fisher	Telephone	2	15/30/60	600	1920	28.7
Sim 2spk	Read	2	57.9	36.1	6434	39.7
Sim 3spk	Read	3	68.5	43.1	7137	43.3
Sim 4spk	Read	4	94.2	59.4	9848	57.0

Datasets

Baseline

- We use the **turn-to-diarize** + multi-stage clustering as our baseline system
- A **global configuration** for near **SOTA** performance on various testing sets



Baseline system:
turn-to-diarize + multi-stage clustering

Metrics

- ASR: Word Error Rate (WER)
- Speaker diarization: Word Diarization Error Rate (WDER)

$$\mathrm{WDER} = \frac{S_{\mathrm{IS}} + C_{\mathrm{IS}}}{S + C} \tag{2}$$

where,

1. S_{IS} is the number of ASR Substitutions with Incorrect Speaker tokens,
2. C_{IS} is the number of Correct ASR words with Incorrect Speaker tokens,
3. S is the number of ASR substitutions,
4. C is the number of Correct ASR words.

L. Shafey *et al*., "Joint Speech Recognition and Speaker Diarization via Sequence Transduction", Interspeech 2019.

Model architecture: Pretrained ASR

- Encoder:
  - 12 conformer layers of 512-dim with funnel pooling
  - causal
- Decoder:
  - 640-dim embedding based
  - using two previous non-blank tokens
- Joint:
  - 640-dim hidden
  - projection to 4096-dim wordpiece model vocabulary

Botros, Rami, et al. "Tied & reduced RNN-T decoder." arXiv preprint arXiv:2109.07513 (2021).

Model architecture: Diarization

- Aux encoder:
  - Input connected to the 5th conformer layer of ASR encoder
  - 9 LSTM layers with 1024-dim hidden and 512-dim output
- Aux joint:
  - 640-dim hidden
  - Output vocabulary: 8 pre-defined speaker tokens

Wang, Weiran, et al. "Multi-output RNN-T joint networks for multi-task learning of ASR and auxiliary tasks." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

Data preprocessing

- To train the model efficiently, we need a **max sequence length** when building batches
- Training data segmented into chunks of 15, 30 and 60 seconds
- We train a **single model** on the mixture of all training data, and evaluate it on all testing sets

Section 5

Results & observations

Taking a first look

We trained a model by mixing all training data together:

- Callhome: better than baseline
- Fisher/AMI: much worse than baseline
- Simulated:
  - Better on 2/3 speakers
  - Worse on 4 speakers

Table 2: *ASR and diarization performance of the baseline and proposed models. WERs (%) are reported with substitution (S), deletion (D) and insertion (I) error rates.*

Testsets	WER (S/D/I)	WDER (%) Baseline	WDER (%) Proposed
Callhome	45.9 (12.8/9.7/23.3)	10.3	7.7
Fisher	20.5 (8.7/10.4/1.4)	3.6	8.0
AMI	29.6 (8.9/19.9/0.8)	8.7	50.0
Sim 2spk	8.1 (6.4/1.0/0.7)	4.2	4.1
Sim 3spk	8.3 (6.5/1.0/0.8)	4.2	3.6
Sim 4spk	8.1 (6.4/1.0/0.7)	4.5	5.1

Sequence length is a huge challenge!

- During training, we need to segment the data with a max sequence length to build batches and train efficiently (15/30/60 seconds)
- An AMI eval utterance can be **30+ minutes** long!
- **The cross-chunk context is lost!!!**

Evaluation on segmented datasets

On the 30s/60s/120s segmented version of the eval sets:

- Callhome&Fisher: better than baseline
- AMI: still worse than baseline

Table 3: *Short-form test WDER (%) on various audio durations.*

Testsets	Short-form Lengths (s)	WDER (%) Baseline	WDER (%) Proposed
Callhome	30	13.6	9.3
	60	9.8	8.9
	120	10.5	8.9
Fisher	30	8.6	3.8
	60	4.8	3.7
	120	4.0	3.7
AMI	30	10.1	9.9
	60	6.7	13.3
	120	8.0	18.8

AMI: break down by number of speakers

On the 30s/60s/120s segmented version of AMI:

- WEEND is better than baseline only for 30-sec and 60-sec, and only when there are at 1~2 speakers

Table 4: *Pre-segmented short-form AMI WDER (%), breakdown by reference number of speakers. For each testset, we compute the WDER for each subset with the same number of ground truth speakers. For the evaluation on 120-sec segments, since there are only 6 single speaker test examples, we do not list these results.*

AMI	Baseline WDER (%)				Proposed WDER (%)			
Lengths	1spk	2spk	3spk	4spk	1spk	2spk	3spk	4spk
30-sec	18.6	10.0	8.8	8.4	1.1	5.8	10.1	15.5
60-sec	10.8	6.3	5.6	6.9	0.8	5.2	12.2	17.1
120-sec	-	6.4	4.4	9.3	-	9.8	15.8	20.8

Why the limitations?

The WEEND model we trained has two major limitations:

- Poor performance on **longform utterances**:
  - We applied **max sequence length** of 60s for efficient training
  - Some testing utterances are much longer than that

- Poor performance on utterances with **more than 2 speakers**:
  - Single model trained on extremely **unbalanced training data**
  - Only AMI has utterances with more than 2 speakers
  - AMI is too small, compared with Fisher

Proof of our explanation

WEEND will be much much worse, if we:

- Remove simulated data
- Only train on segments of 15 seconds

Table 6: *Training data augmentation impact on WDER (%). The second row excludes simulated data from training. The last row further drops 30/60s training segments, i.e. only trained on 15s data.*

Model	CH	CH Short	Fisher	Fisher Short	Sim	AMI Short
Proposed	7.7	9.0	8.0	3.7	4.3	14.0
-Simulated	11.6	9.8	12.3	5.4	22.2	19.0
-30/60s segs	28.8	22.5	22.1	15.7	26.8	26.2

Ablation study: where to hook aux encoder

As an aux task, speaker diarization benefits from both:

- **Acoustic** hints: thus last layer won't work
- **Semantic** hints: thus input layer won't work

Table 5: *Impact of intermediate layer selection on WDER (%). Callhome is abbreviated as CH. Average numbers are reported for simulated and short-form AMI.*

Intermediate Layer Selection	CH	Fisher	Sim	AMI Short
0th Conf layer (features)	23.8	24.5	10.4	22.8
5th Conf layer (proposed)	7.7	8.0	4.3	14.0
12th Conf layer (last)	33.6	37.3	46.9	27.5

Section 6

Conclusion & future work

WEEND: The success

- Allows us to add speaker diarization capability to any off-the-shelf transducer-based ASR model
- No speaker embeddings; no clustering
- Zero regression on ASR tasks
- Outperforms baseline on shortform & 2-speaker cases

De Silva, Sam, Anthony Liu, and L. L. P. Nabarro. "Europe's tough new law on biometrics." Biometric Technology Today 2017.2 (2017): 5-7.

WEEND: The limitations

- Poor performance on longform utterances, due to training constraints
- Poor performance on utterances with more than 2 speakers, due to unbalanced data

But the future is bright!

We believe the **limitations are only transient!**

Training with much longer sequences is possible

- Just think about LLMs

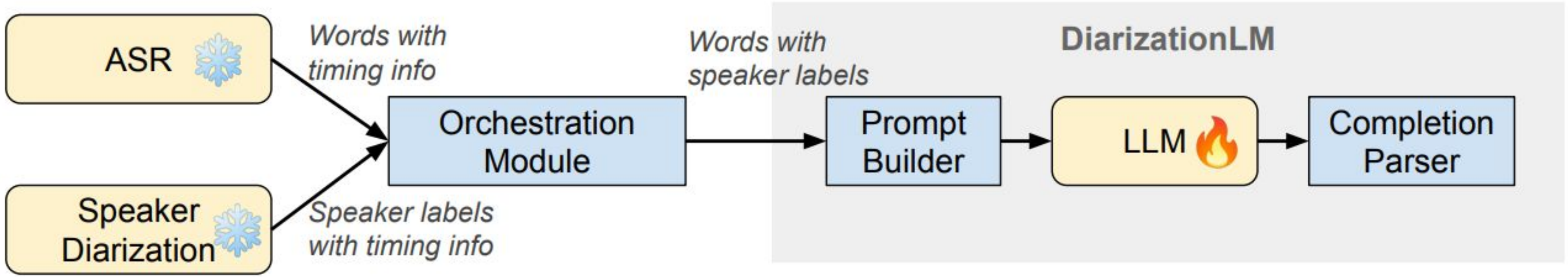
Data issue will ultimately be gone

- Our research is constrained to public datasets, which are extremely unbalanced
- Tons of data are available in the wild (YouTube, bilibili, etc.)

But the future is bright!

And we have **DiarizationLM**!

- Works as a post-processing step of **ANY** ASR + diarization solution
- Please come to our poster session
  - Sep 4th, Wednesday
  - Poster Session: Speaker Diarization 2
  - A4-P4-A
  - Location: Poster Area 1A



Questions?

INTERSPEECH 2024