Paper Number: 1347

Turn-to-Diarize:

Online Speaker Diarization Constrained by Transformer Transducer Speaker Turn Detection

Wei Xia, Han Lu, Quan Wang, Anshuman Tripathi, Yiling Huang, Ignacio Lopez Moreno, Hasim Sak
Google LLC
{ericwxia · luha · quanw}@google.com

wq2012/SpectralCluster

Problem with Supervised Diarization

- Recently, **supervised** speaker diarization systems are becoming very popular, *e.g.* UIS-RNN, discriminative neural clustering, permutation invariant training
- To train such systems, time-annotated speaker labels are required for the training data
- This kind of annotation is extremely expensive and error-prone: single pass annotation of **10 min** of audio takes about **2 hours**!
- Alternatively, annotating speaker turns only adds minimal incremental efforts on top of regular ASR transcribing

```
start: 0.0, end: 1.2, speaker: A
start: 1.3, end: 4.4, speaker: B
start: 6.7, end: 9.4, speaker: A
```

Time-annotated speaker labels

```
good morning <st>
morning how are you <st>
good what about you <st>
```

Speaker turns on top of ASR transcription

System Overview

- Make use of speaker turn annotations that can be acquired at large scale
- A transformer transducer for joint speaker turn detection and ASR
- Compute turn-wise speaker embeddings
- Constrain the unsupervised clustering algorithm with speaker turns

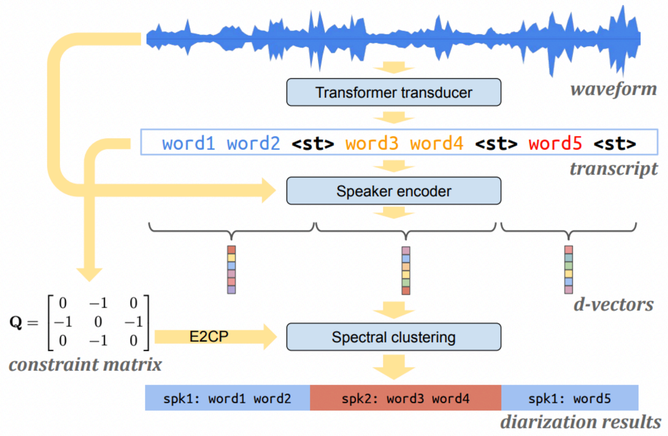


Fig. System architecture of the speaker diarization system.

Speaker Turn Detection

- 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>`)

Table 1. Hyper-parameters of a Transformer block.

Input feature projection	256
Dense layer 1	1024
Dense layer 2	256
Number attention heads	8
Head dimension	64
Dropout ratio	0.1

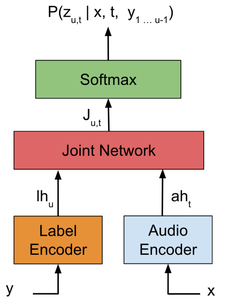


Fig. Transformer transducer.

Turn-to-Diarize

Speaker turn detection:

- If a turn is longer than 6s, insert a **fake turn**
- This compensates for false rejects

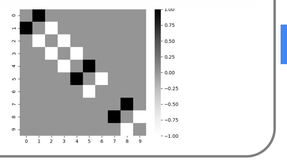
Speaker encoder:

- Reset states at speaker turns, one embedding for each turn
- This makes the sequence much shorter and clustering much cheaper!

Spectral clustering:

- Few modifications from previous work – no Gaussian blur; thresholding with auto-tune; using Laplacian matrix
- Apply a constraint matrix to the affinity matrix

$$\mathbf{Q}_{ij} = \begin{cases} -1, & \text{If } (i,j) \in \text{CL and } c(\texttt{<st>}) > \sigma; \\ +1, & \text{If } (i,j) \in \text{ML}; \\ 0, & \text{Otherwise.} \end{cases}$$



E2CP

$$\mathbf{Q}^* = (1-\alpha)^2 (I - \alpha\bar{\mathbf{A}})^{-1} \mathbf{Z} (I - \alpha\bar{\mathbf{A}})^{-1}$$

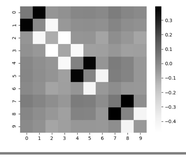


Fig. Initial constraint matrix.
"Cannot-Link" (CL): segmented by true `<st>`;
"Must-Link" (ML): segmented by fake `<st>`.

Fig. Constraint matrix after Exhaustive and Efficient Constraint Propagation (E2CP).

Experimental Results

Datasets:

- Speaker turn detection training set: ~7500 hours of YouTube + public data
- Speaker encoder training set: Vendor collected data (37 locales) + public data
- Diarization eval set: "Outbound" telephone (2 speakers); "Inbound" telephone (2~10 speakers); Callhome American English (eval subset)

Metrics:

- Quality: speaker confusion (Conf.) and diarization error rate (DER)
- Efficiency: floating point operations to process 1s of audio (FLOP/s)

Table 2. Confusion (%), total DER (%) and GFLOP/s on three datasets for different embeddings and methods.

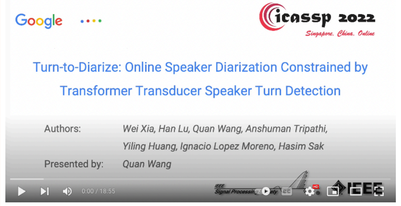
System	Method	Inbound		Outbound		Callhome Eval		GFLOP/s at 10min	GFLOP/s at 1h
		Conf.	DER	Conf.	DER	Conf.	DER		
Dense d-vector	Dense	17.98	22.13	10.66	15.97	5.39	7.76	0.85	36.54
	Dense + Auto-tune	14.09	18.24	9.56	14.88	5.42	7.79	4.76	361.37
Turn-to-diarize	Turn	17.87	19.43	8.41	10.34	8.23	10.08	1.00	1.18
	Turn + E2CP	17.21	18.77	7.94	9.86	3.56	5.41	1.00	1.18
	Turn + Auto-tune	13.83	15.39	7.01	8.93	5.11	6.95	1.02	2.81
	Turn + E2CP + Auto-tune	**13.66**	**15.22**	**6.86**	**8.78**	**3.49**	**5.33**	1.02	2.81

Results:

- "Dense d-vector" is extremely expensive for long-form speech – *eigen-decomposition of a huge Laplacian matrix*
- Turn-to-diarize is more feasible for long-form speech – *speaker turns are sparse, thus Laplacian matrix is very small, even after 1h of audio*
- Auto-tune and speaker turn constraints are critical for turn-to-diarize performance
- Best turn-to-diarize significantly outperforms best "dense d-vector"

More Information

Full lecture is available on YouTube



Clustering algorithm on GitHub
https://github.com/wq2012/SpectralCluster