Frame-wise streaming end-to-end speaker diarization with non-autoregressive self-attention-based attractors

Di Liang, Nian Shao, Xiaofei Li

Westlake University, Hangzhou, China

Introduction

Topic Streaming Speaker Diarization: answering "who speaks when" in streaming applications

Method Deep-learning based streaming speaker diarization with frame-wise attractors

- Frame-wisely detect a flexible number of speakers and extract/update their corresponding attractors.
- A look-ahead mechanism allows leveraging some future frames.

Network

- A causal speaker embedding encoder by masked self-attention module.
- Look-ahead with 1-dimensional convolution.
- An online attractor decoder to extract frame-wise attractors.

Contributions

- Excellent speaker diarization performance on both simulated dataset and real-world CALLHOME data.
- Computational cost is lower compared to block-wise online methods.

Method

Formulation: a multi-class detection task

- $X = (\boldsymbol{x}_t \in \mathbb{R}^F | t = 1, \ldots, T)$: LogMel feature vector sequence
- $Y = (\boldsymbol{y}_t \in \{0,1\}^S | t = 1, \ldots, T)$: speaker label sequence
- F: dimension of feature vector
- S: number of speakers

Network Architecture

The basic strategy is to design

✓ **a causal speaker embedding encoder**: masked self-attention and Conv1D along time dimension.

✓ **an online attractor decoder**: extract attractors frame-wisely and is realized with a non-autoregressive self-attention network.

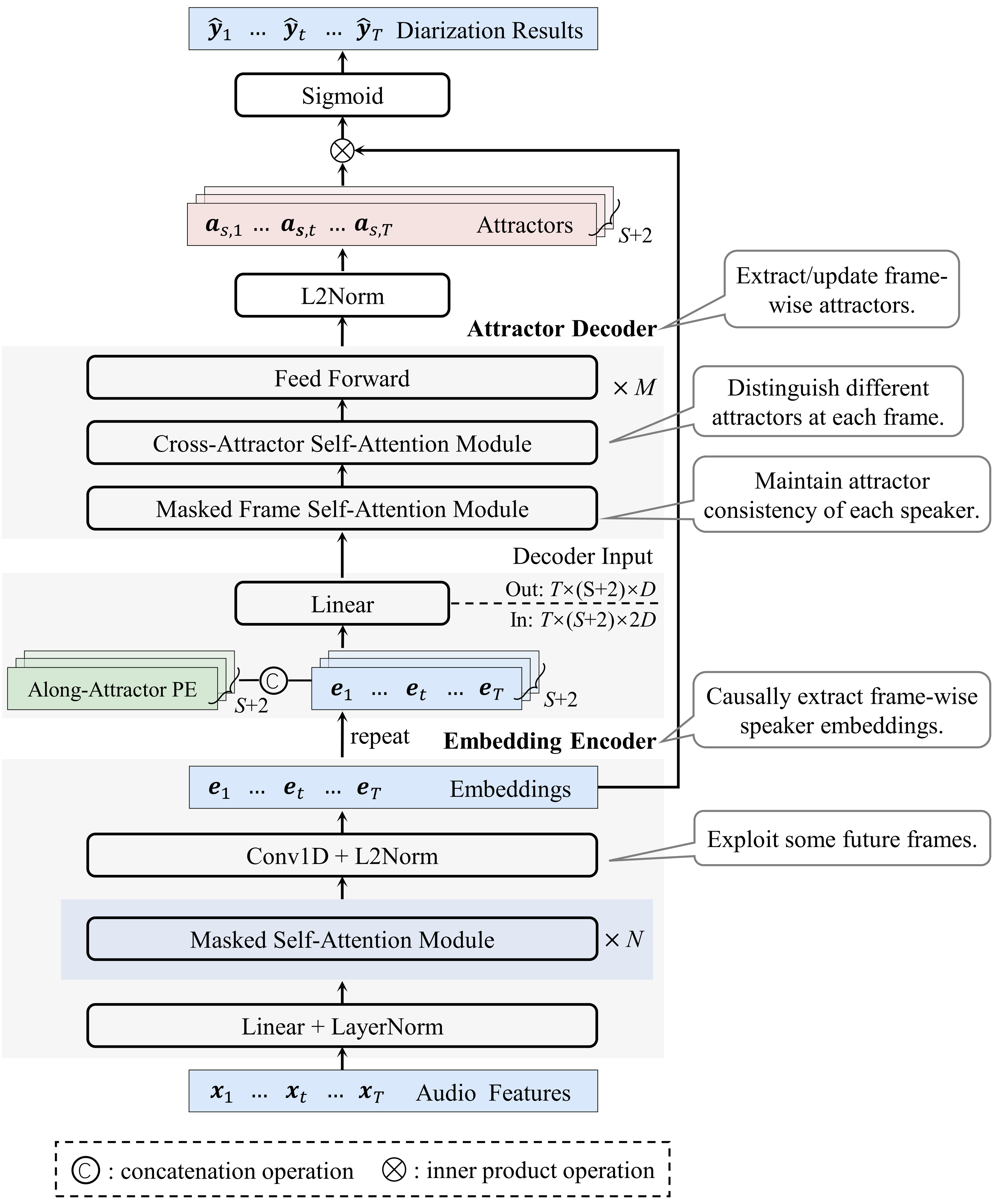

Figure: Architecture of the proposed FS-EEND system.

Experiments

Dataset

- Simulated data: Extract single-speaker utterances using VAD and generate (1,2,3,4)-speaker mixtures with noises and reverberation.
  - Data source: Switchboard Cellular (Part 1 and 2) & 2005-2008 NIST Speaker Recognition Evaluation (SRE)
  - Speakers: 3244, 405, and 405 for training, validation, and test set
- CALLHOME data: Telephone call voice data set

Metrics

- Diarization error rate (DER)

Results

- Results on simulated data

Table: DERs (%) and RTF on simulated data.

Methods	RTF	latency (s)	Number of speakers 1	2	3	4
Offline EEND-EDA	0.006	-	0.4	4.0	9.9	14.1
EEND-EDA+FLEX-STB	0.028	10	0.7	4.7	13.0	17.1
EEND-EDA+FLEX-STB	0.223	1	1.9	6.7	15.1	19.6
FS-EEND (prop.)	0.026	1	0.6	5.1	11.1	15.8

- Results on CALLHOME data

Table: DERs (%) on CALLHOME data.

Methods	latency (s)	Number of speakers 2	3	4
Offline EEND-EDA	-	7.7	13.7	22.4
BW-EEND-EDA	10	11.8	18.3	25.9
EEND-EDA+FLEX-STB	10	9.6	14.4	22.0
EEND-EDA+FLEX-STB	1	13.0	16.4	23.6
FS-EEND (prop.)	1	10.1	14.6	21.2
EEND-EDA+FLEX-STB+VCT	1	11.1	16.0	21.7
FS-EEND+VCT (prop.)	1	9.4	14.0	20.9

- t-SNE visualization of speaker embeddings

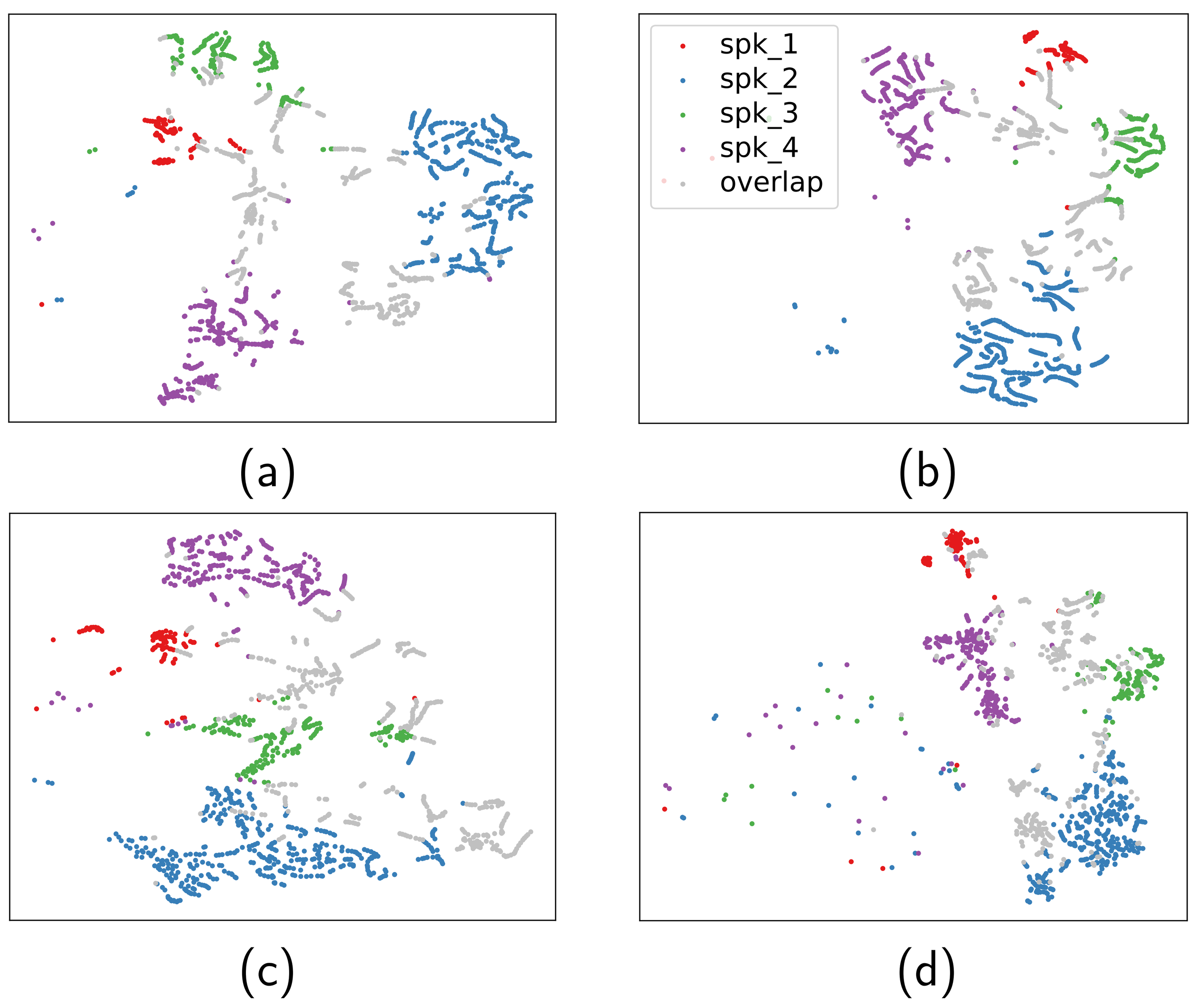

Figure: t-SNE visualization of embeddings in 2-dimensional space. (a) the proposed FS-EEND, (b) without L2-normalization, (c) without embedding similarity loss and (d) without look-ahead.

Conclusions

- FS-EEND processes audio stream and performs diarization frame by frame, with a causal embedding encoder and an online attractor decoder.
- FS-EEND shows superiority in diarization performance, system latency, and computational complexity.

Code[a]

[a] https://github.com/Audio-WestlakeU/FS-EEND