FANO 有光科技 Labs

End-to-End Speaker Diarization System for the Third DIHARD Challenge

Tsun-Yat Leung, Lahiru Samarakoon

Fano Labs, Hong Kong
ty.leung@fano.ai, lahiru@fano.ai

Outline

Introduction

Proposed System

Datasets

Results

Introduction

Based on “self-attentive end-to-end diarization model with encoder-decoder based attractors (EDA-EEND)”

[Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., Nagamatsu, K. (2020). End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors.]

Focus on DIHARD III track 2

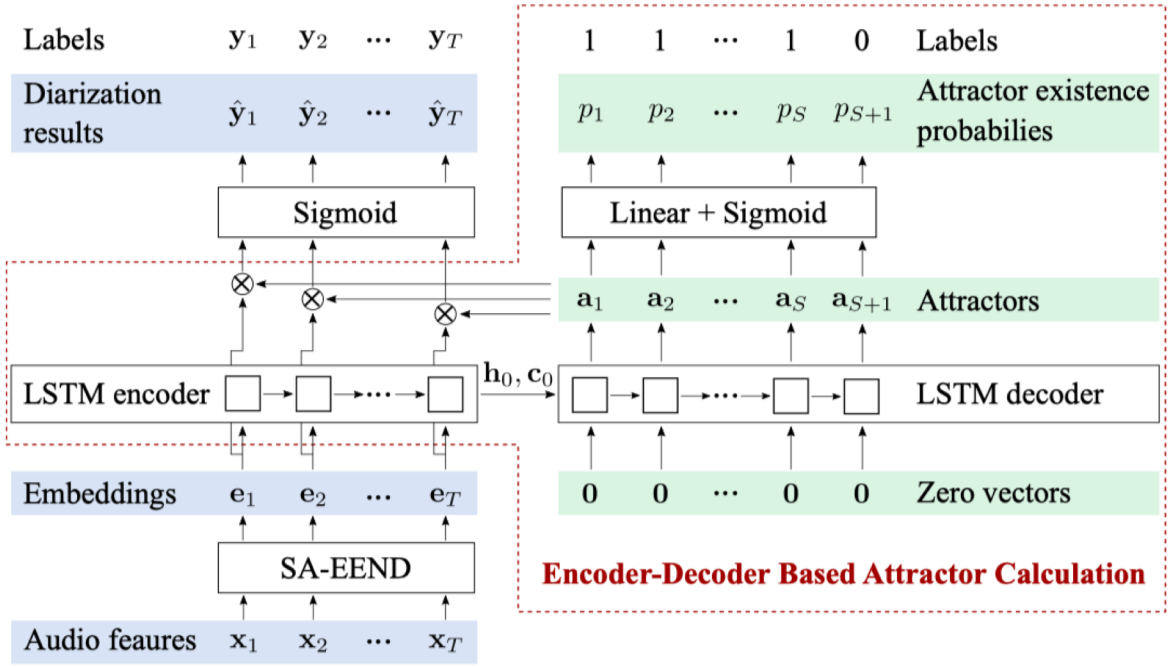


Figure 1: *SA-EEND with encoder-decoder based attractor calculation.*

Proposed System

FANO 有光科技 Labs

5 Modifications on EDA-EEND

- Conformer Encoders
- Convolutional Upsampling
- Attractor Calculation with Attentions
- Additive Margin Penalty
- Chunk Shuffling

Conformer

- Transformers + Convolution networks

- Capturing fine-grained local features

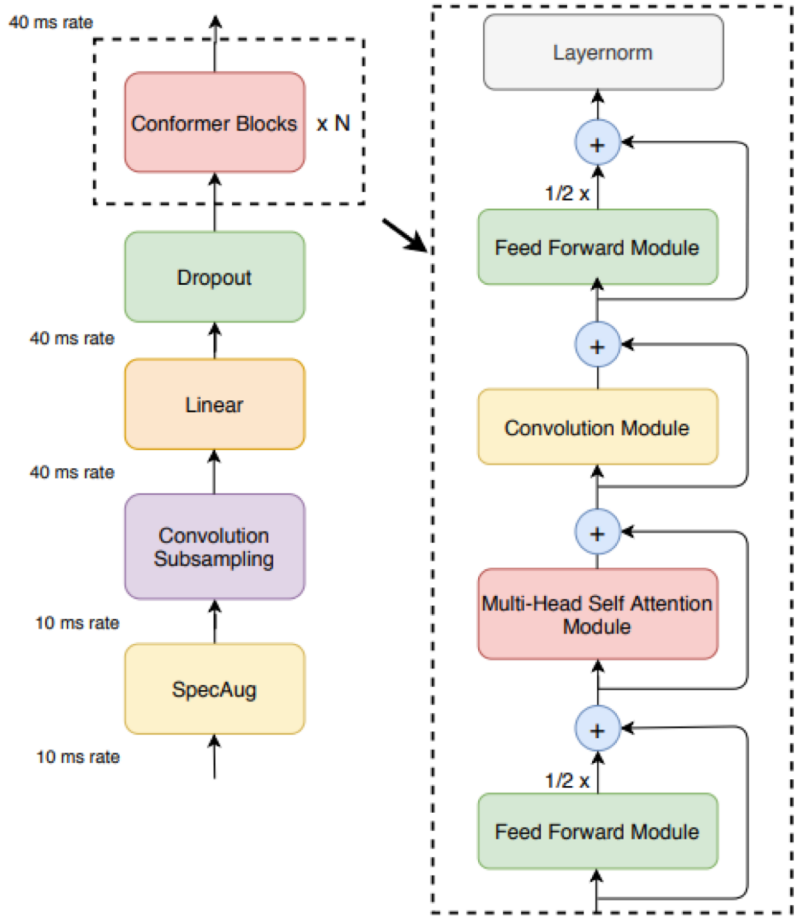


Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*

Gulati, Anmol, et al. "Conformer: Convolution-augmented Transformer for Speech Recognition." (2020).

Convolutional Upsampling

FANO 有光科技 Labs

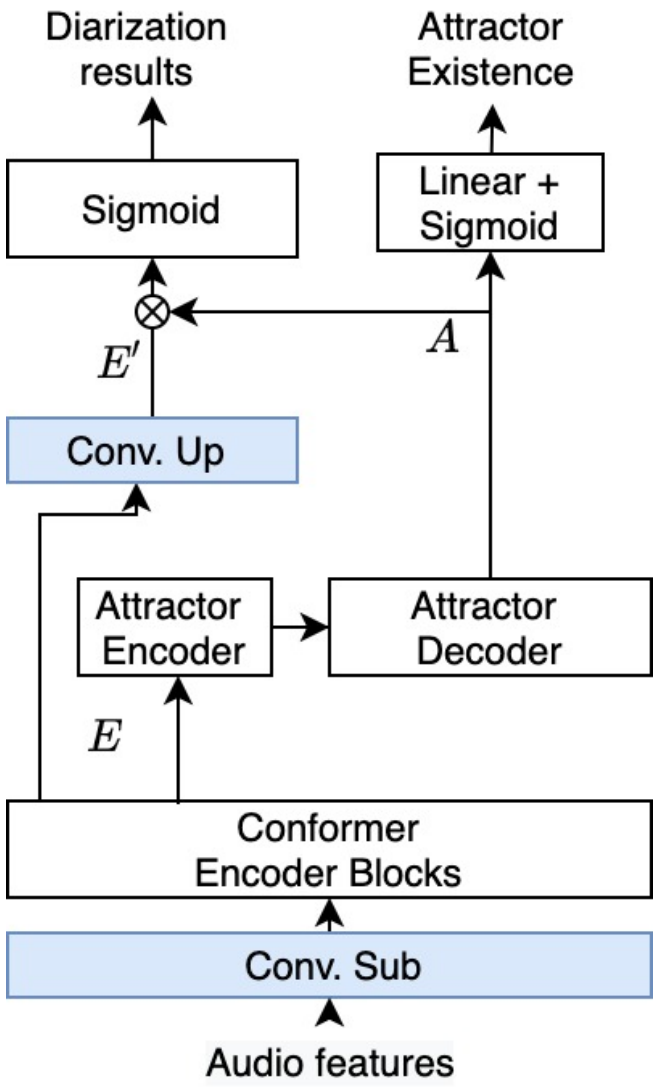
Observations:

- Each frame in EDA-EEND is 0.1s

- No collar in the evaluation

- Result with Low resolution => Increase in DER

Convolutional Upsampling (Cont.)



Attractor Calculation with Attentions

FANO 有光科技 Labs

Problems/Observations:

- Long sequence of Embeddings

- Information of the attractors passed by the Last timestamp of the Encoder Outputs.

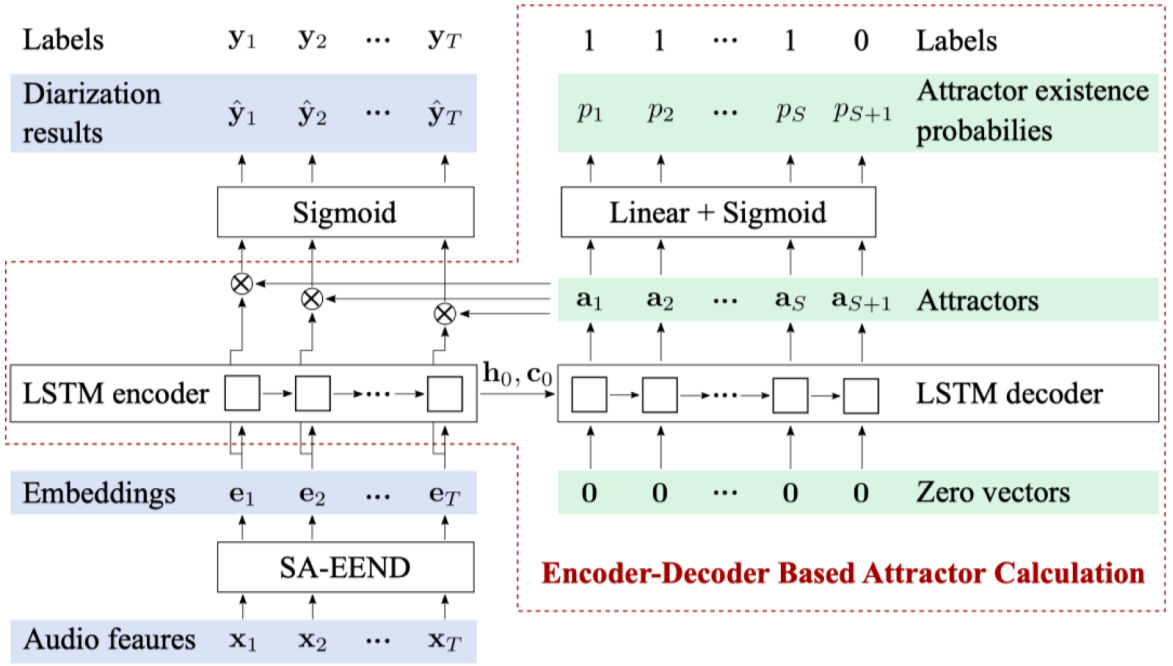
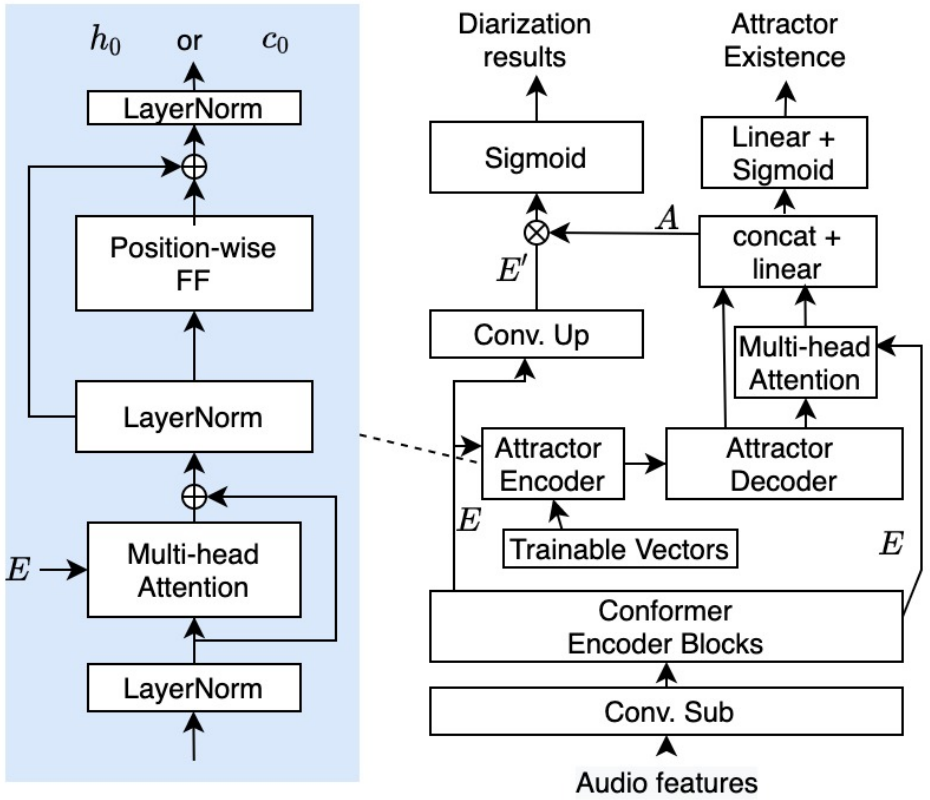


Figure 1: *SA-EEND with encoder-decoder based attractor calculation.*

Attractor Calculation with Attentions (Cont.)

Changes:

- Multi-head attentions as pooling mechanism to initialize h_0 and c_0.
- Global Attentional mechanism on the Decoder.



Additive Margin Penalty

FANO 有光科技 Labs

Put Additive Margin Penalty into speaker diarization result calculation:

- All attractors are normalized

- Using PIT to obtain correct permutation of speaker labels

- Posterior probability $\hat{y}_{t,s}$ of speaker s at time t:

$$\hat{y}_{t,s} = sigmoid(\gamma(\mathbf{e}_t\mathbf{a}_s - y_{t,s}m + (1 - y_{t,s})m))$$

$y_{t,s} \in \{0, 1\}$ is the label of speaker s at t, $\mathbf{e}_t$ is the embedding at time t, $\mathbf{a}_s$ is the attractor of speaker s, γ is the scale factor, m is the additive margin value.

Chunk Shuffling

FANO 有光科技 Labs

During training, each training sample is a 50 seconds audio.

To increase the combinations of different audio segments:

- divide the original recording into chunks of 10 seconds

- shuffle chunks with a probability of 0.5.

Datasets

FANO Labs 有光科技

Table: Training and validation datasets

Dataset	#Mixtures
Pretraining Training Set	
Librispeech Simulated (β=2, 2, 4, 6)	400,000
Pretraining Validation Set	
Librispeech Simulated (β=2, 2, 4, 6)	2000
Fine-Tuning Training Set	
VoxConverse Development	216
DIHARD III Development (Training)	203
DIHARD II Development Clinical	24
Fine-Tuning Validation Set	
DIHARD III Development (Validation)	51

Results of Conformer and Resolution Exp.

FANO 有光科技 Labs

Table: Track 2 result of conformer and resolution experiments. "Val" refers to our fine-tuning validation set.

Part	Conformer	Deep	Conv. Down & Up	DER (%)		JER (%)	
				Val	Eval	Val	Eval
core	No	No	No	28.29	30.15	54.51	53.20
core	Yes	No	No	27.18	29.03	51.94	**50.86**
core	Yes	Yes	No	25.95	29.05	53.32	52.65
core	Yes	Yes	Yes	**25.06**	**27.90**	**51.85**	52.20
full	No	No	No	26.58	25.70	48.60	46.47
full	Yes	No	No	25.21	24.64	45.82	**44.38**
full	Yes	Yes	No	24.25	24.69	47.28	45.84
full	Yes	Yes	Yes	**22.48**	**23.05**	**45.25**	44.93

"Deep" means that the encoder has 7 layers and hidden unit dims = 128 instead

Results of attractor with attentions and additive margin penalty

FANO 有光科技 Labs

Table: Track 2 result of attractor with attentions and additive margin penalty.

Part	Attractor with Attention	Additive Margin Penalty	DER (%)		JER (%)	
			Val	Eval	Val	Eval
core	No	No	25.06	27.90	53.32	52.20
core	Yes	No	24.05	**26.08**	52.01	51.73
core	Yes	Yes	**22.66**	26.12	**51.43**	**50.87**
full	No	No	22.48	23.05	47.28	44.93
full	Yes	No	22.25	**21.65**	45.75	44.35
full	Yes	Yes	**21.07**	21.70	**45.17**	**43.86**

Results of chunk shuffling and additional training

FANO Labs 有光科技

Table: Track 2 result of chunk shuffling and additional training. "Dev" refers to original DIHARD III development set, and "Larger Epoch" means that the pretrained model is trained with more number of epochs.

Part	Chunk Shuffling	Larger Epoch	DER (%)			JER (%)		
			Val	Dev	Eval	Val	Dev	Eval
core	Yes	No	22.32	**18.33**	24.72	51.44	**42.33**	49.75
core	Yes	Yes	**21.85**	18.64	**23.86**	**50.61**	43.03	**48.85**
full	Yes	No	20.92	**16.36**	20.72	45.21	**36.69**	42.86
full	Yes	Yes	**20.04**	16.53	**20.05**	**44.01**	37.21	**42.06**


THANK YOU!
FANO
有光科技
Labs