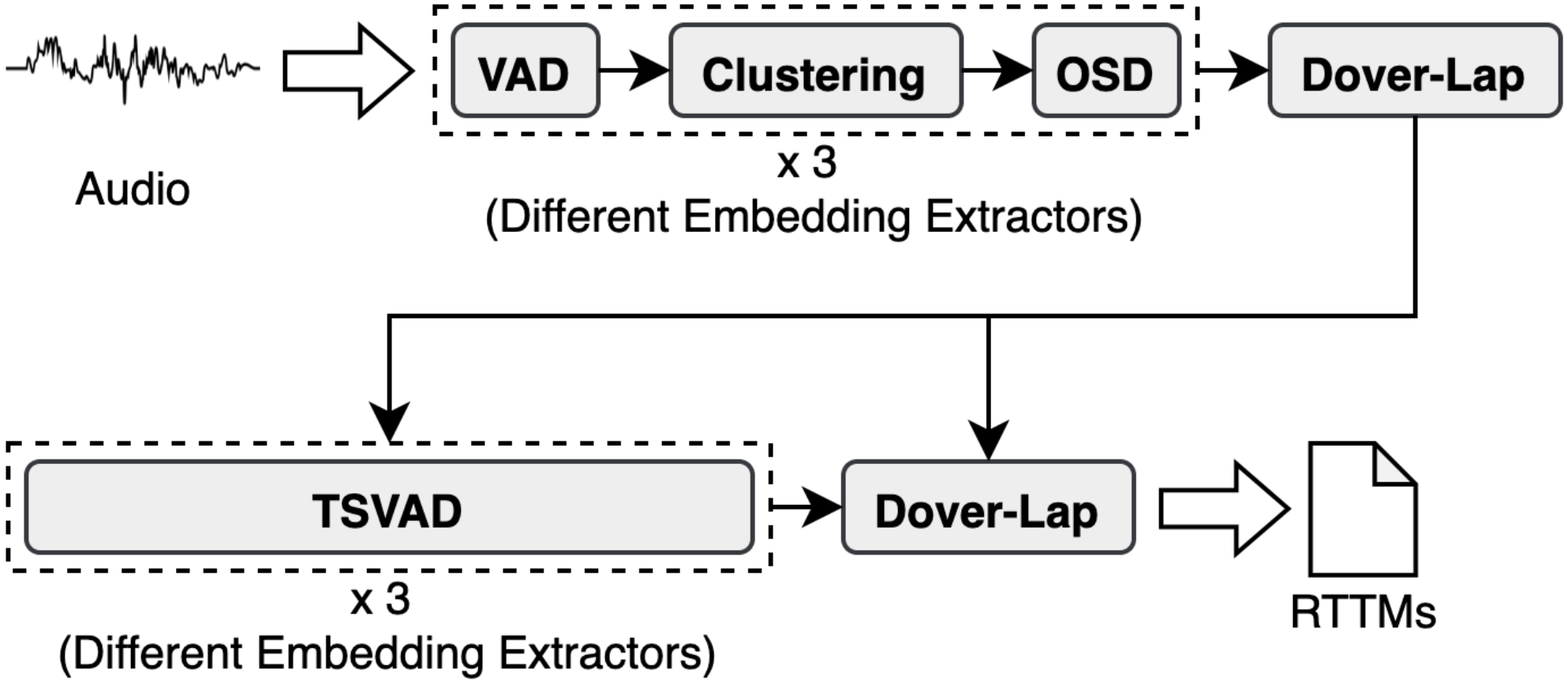
The DKU-MSXF Diarization System for the VoxCeleb Speaker Recognition Challenge 2023

Ming Cheng[1], *Weiqing Wang*[1], *Xiaoyi Qin*[1], *Yuke Lin*[1], *Ning Jiang*[2], *Guoqing Zhao*[2], *Ming Li*[1]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China

[2]Mashang Consumer Finance Co., Ltd.

System Overview



VAD & OSD

- **Model 1:** Conformer Encoder + FC Layer
- **Model 2:** ResNet34 + Spatial Average Pooling + FC Layer
- **Model 3:** ECAPA-TDNN + FC Layer

Table 1: False alarm (FA) and miss detection (MISS) rates of different VAD and OSD models on the VoxConverse test set.

Task	Model	FA (%)	MI (%)	Total (%)
VAD	Conformer	2.84	1.09	3.93
	ResNet34	3.20	1.02	4.22
	ECAPA-TDNN	2.70	1.51	4.21
	Fusion	2.83	1.14	3.97
OSD	Conformer	0.59	1.41	2.00
	ResNet34	0.54	1.51	2.05
	ECAPA-TDNN	0.52	1.45	1.97
	Fusion	0.44	1.45	1.89

Speaker Embedding

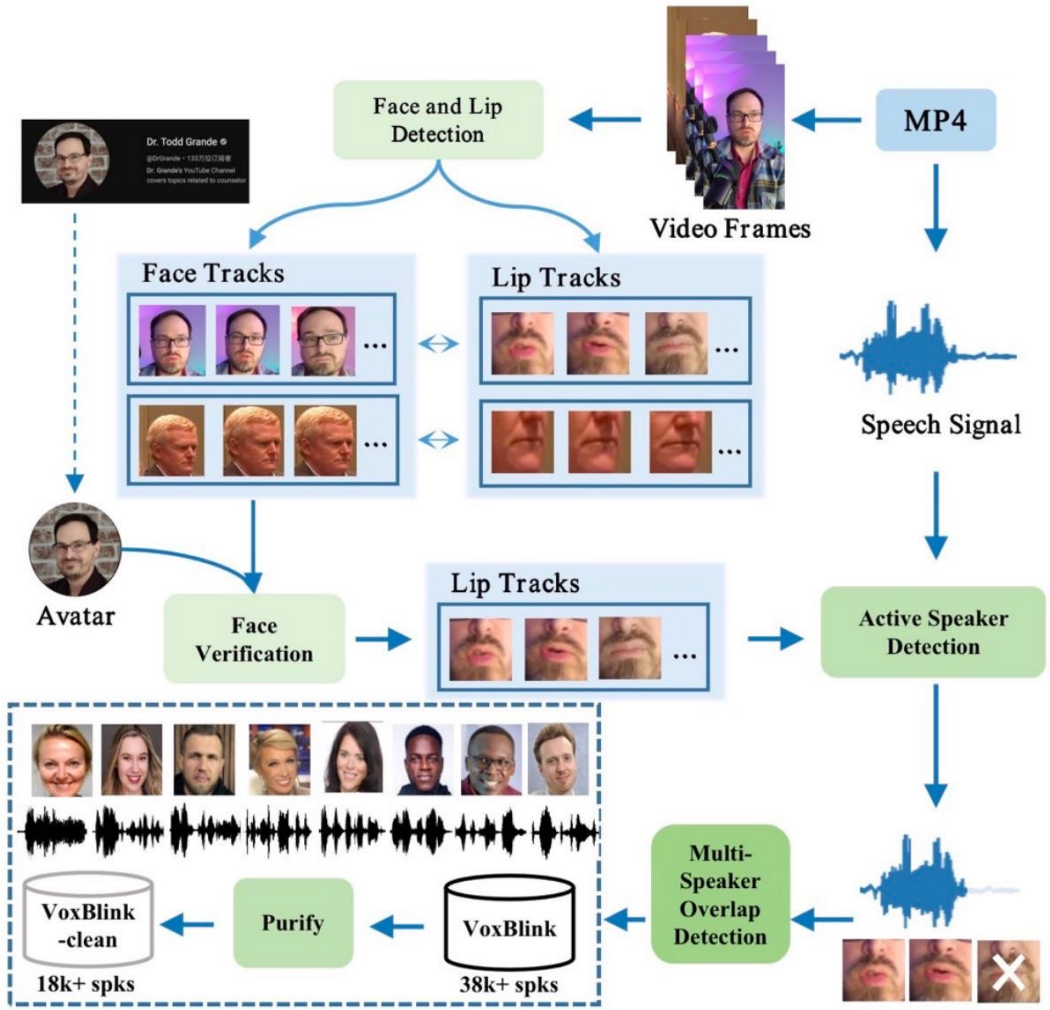
- **Model 1:** SimAM-ResNet34 + Statistics Pooling
- **Model 2:** ResNet101 + Attentive Statistics Pooling
- **Model 3:** SimAM-ResNet100 + Attentive Statistics Pooling

  + Linear (out_dim=256) + ArcFace Classifier

Table 2: Equal error rates (EERs) of different speaker embedding models on the Vox-O trial.

#	Model	Training Data	EER (%)
Spk1	SimAM-ResNet34+SP	Vox2	0.81
Spk2	ResNet101+ASP	Vox2	0.49
Spk3	SimAM-ResNet100+ASP	Vox2+VoxBlink	0.44

Collection of VoxBlink Dataset



- 18k+ identities
- Purified by face verification
- Multimodal Data

Dataset	VoxBlink	VoxBlink-clean
# of SPKs	38,067	18,381
# of videos	372,091	241,172
# of hours	2,135	1,670
# of utterances	1,455,237	1,028,106
Avg # of videos per SPK	9.77	13.12
Avg # of utterances per SPK	38.23	55.93
Avg # of duration per utterance (s)	5.28	4.87
Avg # of video recording intervals (days)	39.72	34.55
Avg # of video recording span (days)	440.07	441.85

The dataset will be released soon: https://arxiv.org/abs/2308.07056

Clustering-based Diarization

AHC for segmentation

- Uniformly segment speech with a length of 1.28s and shift of 0.32s
- Iteratively merge two closest consecutive segments with the largest cosine similarity until the preset threshold is reached

AHC for grouping speaker identities

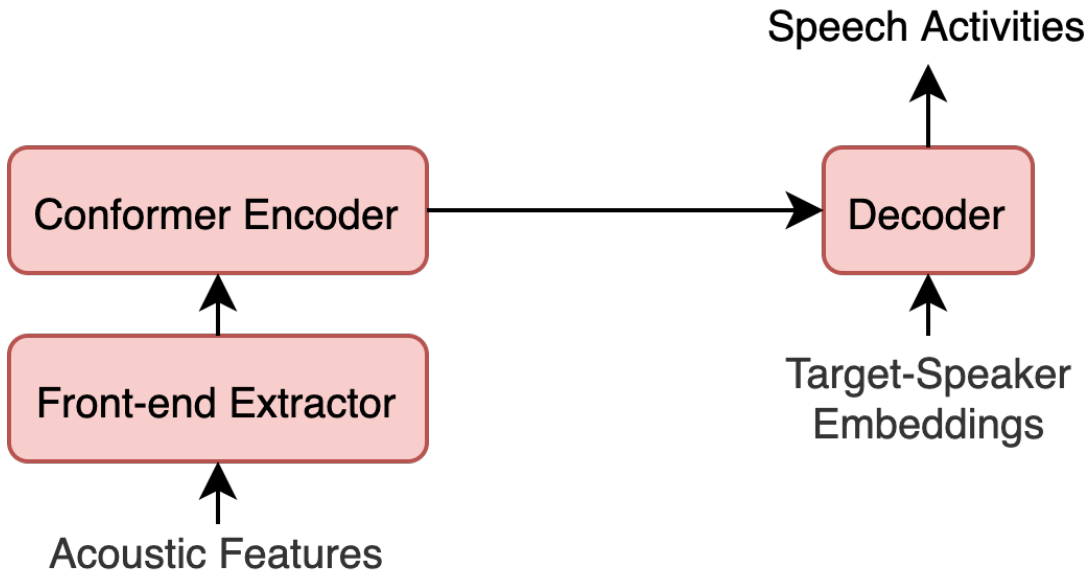
- Perform a plain AHC on the similarity matrix with a relatively high stop threshold to obtain clusters
- Split clusters into “long clusters” and “short clusters” by the total duration in each cluster
- Assign each short cluster to the closest long cluster, and some short clusters are treated as new speakers if not matching any long clusters.
- Assign overlapped speech regions to the two closest speakers

—> Using speaker embedding models in Table 2 as different front-end extractors, we train three sub-models.

TSVAD-based Diarization

Seq2Seq-TSVAD[1]

- Front-end Extractor: initialized by the speaker embedding model same as the one for target-speaker embeddings
- Encoder & Decoder: 6 layers with 512-dim/8-head attention



[1] Cheng, Ming, et al. "Target-speaker voice activity detection via sequence-to-sequence prediction." *In Proc. ICASSP*. IEEE, 2023.

Seq2Seq-TSVAD

Training

BCE loss & Adam optimizer with a learning rate of 1e-4 and warm-up of 2,000 iterations

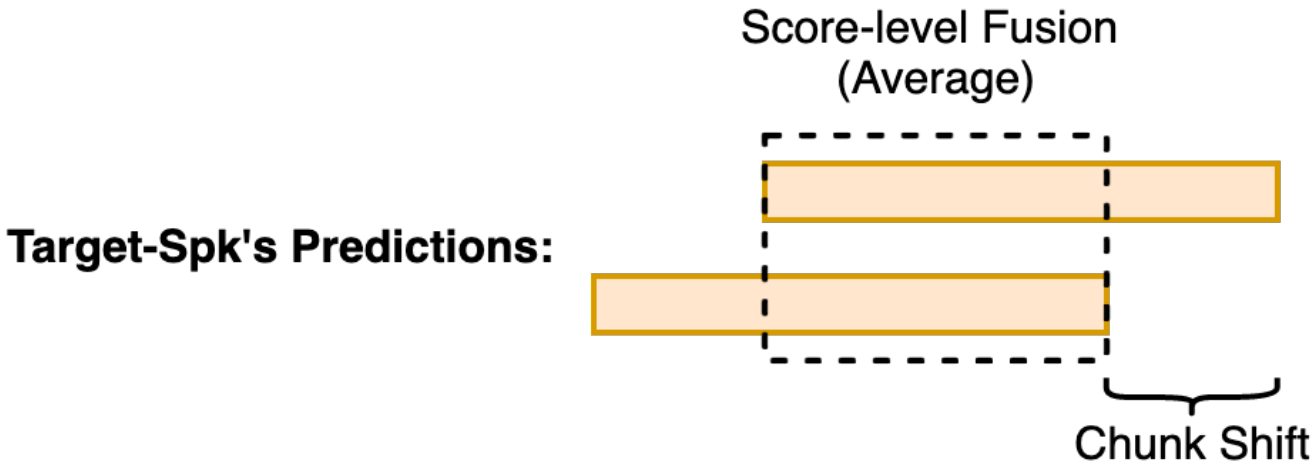
- First, the model with a frozen front-end extractor is trained on simulated data until back-end convergence.
- Second, all model parameters are unfrozen to train on both simulated and real data with learning rate decay.
- Third, all data simulation and augmentation are removed. The first 186 samples of the VoxConverse test set are mixed into finetuning and the last 46 samples are leaved as validation, namely the VoxConverse test46 set.

—> Using speaker embedding models in Table 2 as different front-end extractors, we train three sub-models.

Seq2Seq-TSVAD

Inference

- A clustering-based diarization is required first to extract speaker profiles
- Each test audio is cut into fixed-length chunks with a stride of 1s and fed into the TSVAD model with extracted speaker profiles.
- Chunked predictions are stitched by averaging the overlapped predicted regions, which can also be viewed as a score-level fusion.



Experimental Results

Table 3: Diarization error rates (DERs) of different systems on the VoxConverse test set, VoxConverse test46 set, and VoxSRC-23 challenge test set.

#	Method	DER (%)		
		VoxConverse Test	VoxConverse Test46	VoxSRC-23 Challenge Test
1	Ahc1	4.83	4.14	5.51
2	Ahc2	4.49	3.92	5.32
3	Ahc3	4.55	3.92	5.36
4	Dover-Lap (#1-3)	-	3.81	5.19
5	+ TSVAD with Spk1	-	2.85	4.49
6	+ TSVAD with Spk2	-	2.93	4.57
7	+ TSVAD with Spk3	-	2.91	4.53
8	Dover-Lap (#4-7)	-	2.73	4.30

—> We also test the combination of “Ahc1 and TSVAD-with-Spk1”, which obtains the single-system DER of 4.63% without any Dover-Lap fusion.

Conclusions

Improved Key Points

- Diverse Speaker Embedding models for AHC
- New TSVAD methods with good neural network training

DER reduction from single-system diarization to multi-system fusion

- AHC: 5.32% --> 5.19%
- Seq2Seq-TSVAD: 4.49% --> 4.30%