The Krisp Diarization system for the VoxCeleb Speaker Recognition Challenge 2023

Davit Karamyan[1,2], **Grigor Kirakosyan**[2,3]

[1] *Russian-Armenian University, Yerevan*
[2] *Krisp.ai, Yerevan*
[3] *Institute of Mathematics of NAS RA, Yerevan*

Outline

Voice Activity Detection

- #1 GRU-based: 4 layers of GRU + layer norm + linear
- #2 NC-based: Noise cancellation + energy threshold + post-processing
- #3 Conformer ASR[1]: Word timestamps + post-processing
- #4 Pyannote: Pretrained *Pyannote 2.1*[2]

[1] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_medium
[2] https://huggingface.co/pyannote/segmentation

Voice Activity Detection

- #1 GRU-based: 4 layers of GRU + layer norm + linear
- #2 NC-based: Noise cancellation + energy threshold + post-processing
- #3 Conformer ASR: Word timestamps + post-processing
- #4 Pyannote: Pretrained *Pyannote 2.1*

Table: Detection Error Rate of the VAD model on Voxconverse test set.

#Model	FA	MISS	Detection Error
#1	2.59%	1.40%	3.99%
#2	2.83%	2.09%	4.92%
#3	3.04%	1.74%	4.79%
#4	2.01%	1.19%	3.20%
Fusion	2.02%	0.82%	2.84%

Speaker Embedding

- Pretrained models: TitaNet-Large[3], RawNet3[4] and ECAPA-TDNN[5]
- For *noise robustness* we finetune Titanet-Small with Teacher-Student method[6] on Voxceleb1 and Voxceleb2 dev sets

Table: Equal Error Rate values for different embedding extraction models

Embedding	EER	Training Datasets
TitaNet-Large	0.68% Vox1-Clean	Voxceleb1+Voxceleb2, Fisher, Switchboard, Librispeech
TitaNet-Small*	1.03% Vox1-Clean	Voxceleb1+Voxceleb2
RawNet3	0.89% Vox1-O	Voxceleb1+Voxceleb2
ECAPA-TDNN	0.80% Vox1-Clean	Voxceleb1+Voxceleb2

[3] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/titanet_large
[4] https://huggingface.co/jungjee/RawNet3
[5] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
[6] http://93.187.165.2/index.php/mpcs/article/view/789

Teacher-Student

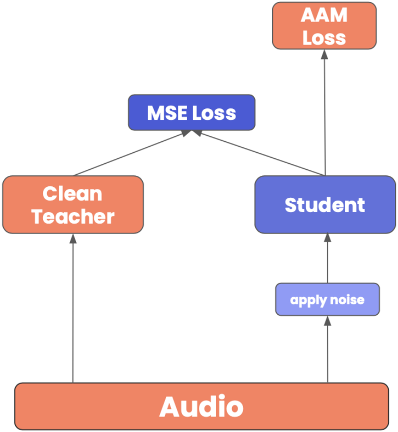

Figure: The flow chart of teacher-student method for improving noise robustness, where the teacher is a pretrained TitaNet-Small model

Spectral Clustering (SC)

- **Multi-scale segmentation**: Weighted sum of affinity matrices for different scales.
- **Affinity Refinement**: Row-wise thresholding + symmetrization + diffusion
- **Maximal eigen-gap** approach to detect number of speakers
- **K-means++** on spectral embeddings

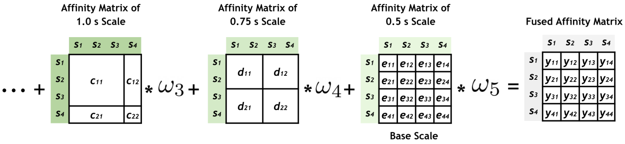

Figure: Multi-scale segmentation scheme

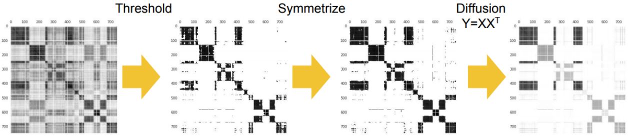

Figure: Refinement operations on the affinity matrix

Agglomerative Hierarchical Clustering

Similar to DKU-SMIIP system[7] in VoxSRC 2022

- Extract speaker embeddings from uniformly segmented speech regions
- Refine embeddings through *dimensionality reduction* and *affinity aggregation (AA)* techniques
- Merge consecutive segments into a longer one if the distance is greater than a segment threshold
- Perform a plain AHC on the refined embeddings with a relatively high stop threshold to obtain the clusters with high confidence
- 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

[7] Wang, W., Qin, X., Cheng, M., Zhang, Y., Wang, K. & Li, M. The dku-smiip diarization system for the voxceleb speaker recognition challenge 2022. *Voxsrc Workshop*. (2022)

Overlap Speech Detection

- *Pyannote overlap speech detection* pipeline[8]
- After an overlapped region is detected, we replace the label with the two closest speakers near this region

[8] `https://huggingface.co/pyannote/overlapped-speech-detection`

Results

Table: The performance of different speaker diarization systems.

N	System	Window [s]	Shift [s]	Voxconverse Test	VoxSRC-23 Test	
				DER[%]	DER[%]	JER[%]
	VGG baseline	-	-	-	8.68	26.71
#1	Pyannote VoxSRC22	-	-	5.89	7.33	33.8
#2	Pyannote VoxSRC22+AA	-	-	5.30	-	-
#3	TitaNet-Large-SC	1.0	0.75	6.00	-	-
#4	TitaNet-Large-SC	2.0	1.0	5.59	-	-
#5	TitaNet-Large-SC	[2.0, 1.5, 0.75]	[1, 0.5, 0.25]	5.25	-	-
#6	ECAPA-TDNN-SC	1.0	0.75	6.05	-	-
#7	ECAPA-TDNN-SC	2.0	1.0	5.71	-	-
#8	ECAPA-TDNN-SC	[2, 1.5, 0.75]	[1, 0.5, 0.25]	5.38	-	-
#9	TitaNet-Small-SC	1.5	0.5	5.23	-	-
#10	TitaNet-Large-AHC	1.5	0.5	5.41	-	-
#11	ECAPA-TDNN-AHC	1.5	0.5	5.38	-	-
#12	RawNet3-AHC	1.5	0.75	5.32	-	-
	Fusion(3+4+5+6+7+8)+OSD	-	-	4.80	6.35	33.71
	Fusion(2+3+4+5+6+7+8)+OSD	-	-	4.76	5.98	31.56
	Fusion(2+5+8+9+10+11+12)+OSD	-	-	**4.39**	**4.71**	29.83

References

- Bredin, H. & Laurent, A. End-to-end speaker segmentation for overlap-aware resegmentation. *Proc. Interspeech 2021*. (2021,8)
- Von Luxburg, U. A tutorial on spectral clustering. *Statistics And Computing*. **17**, 395-416 (2007)
- Park, T., Kumar, M. & Narayanan, S. Multi-scale speaker diarization with neural affinity score fusion. *ICASSP 2021-2021 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 7173-7177 (2021)
- Wang, Q., Downey, C., Wan, L., Mansfield, P. & Moreno, I. Speaker diarization with LSTM. *2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 5239-5243 (2018)
- Wang, W., Qin, X., Cheng, M., Zhang, Y., Wang, K. & Li, M. The dku-smiip diarization system for the voxceleb speaker recognition challenge 2022. *Voxsrc Workshop*. (2022)
- Karamyan, D., Kirakosyan, G. & Harutyunyan, S. Making Speaker Diarization System Noise Tolerant. *Mathematical Problems Of Computer Science*. **59** pp. 57-68 (2023)

References

- D.Raj, P.Garcia, Z.Huang, S.Watanabe, D.Povey, A.Stolcke & S.Khudanpur DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs. *2021 IEEE Spoken Language Technology Workshop (SLT)*. (2021)
- Kwon, Y., Jung, J., Heo, H., Kim, Y., Lee, B. & Chung, J. Adapting speaker embeddings for speaker diarisation. *ArXiv Preprint ArXiv:2104.02879*. (2021)
- Koluguri, N., Park, T. & Ginsburg, B. TitaNet: Neural Model for speaker representation with 1D Depth-wise separable convolutions and global context. *ICASSP 2022-2022 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 8102-8106 (2022)
- Jung, J., Kim, Y., Heo, H., Lee, B., Kwon, Y. & Chung, J. Pushing the limits of raw waveform speaker recognition. *Proc. Interspeech*. (2022)
- Desplanques, B., Thienpondt, J. & Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. *Interspeech 2020*. pp. 3830-3834 (2020)