Spoofing-aware Speaker Verification Robust against Domain and Channel Mismatches

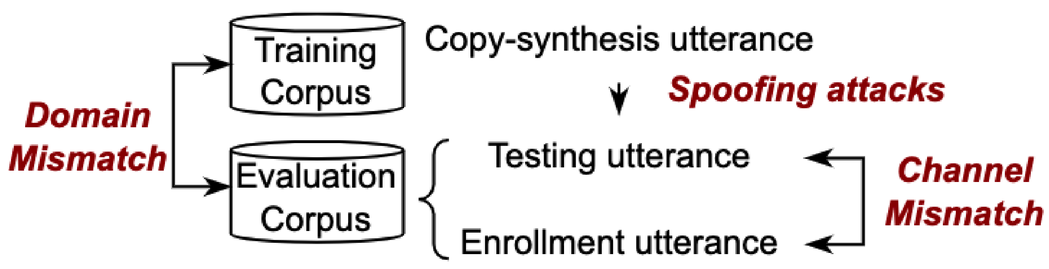
P6-16-SLR (#10)

Chang Zeng[1], Xiaoxiao Miao[2], Xin Wang[1], Erica Cooper[3], Junichi Yamagishi[1]

[1]National Institute of Informatics, [2]Singapore Institute of Technology, [3]National Institute of Information and Communications Technology

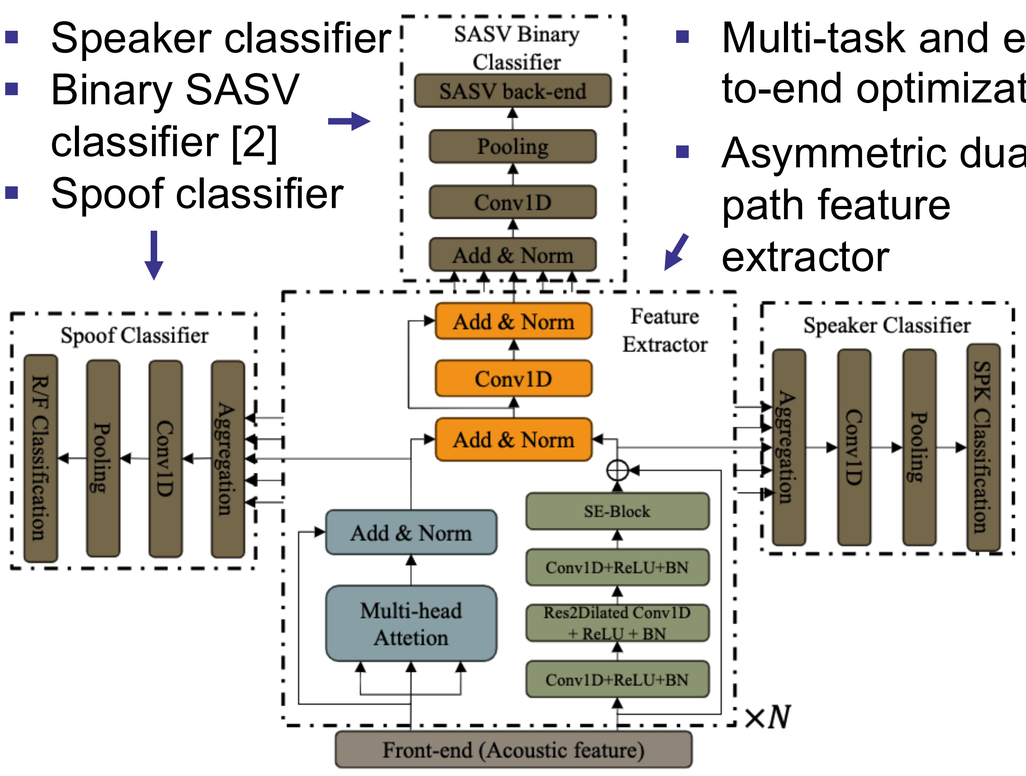
Introduction

- Issue: speaker verification faces spoofing, channel, and domain mismatches.
- Existing methods:
  - Address the mismatches separately.
  - Integrated solutions are rare and underperform in complex scenarios.
- We propose an integrated framework using *meta-learning*, *spoofing attack simulation*, and *multi-task learning* to enhance robustness against the mismatches.



Proposed method

- Model architecture
  - Speaker classifier
  - Binary SASV classifier [2]
  - Spoof classifier
  - Multi-task and end-to-end optimization
  - Asymmetric dual-path feature extractor



- Learning paradigm [1]
  - Simulating domain mismatch using meta-learning
  - Inner loop: training on spoof and channel mismatch scenarios
  - Outer loop: optimizing across unseen domains

Experiment and Results

- Dataset
  - Created CNComplex, a challenging ASV dataset
  - Combined CNCeleb and CNSpoof [1] datasets
  - Evaluate three simultaneous threats: spoofing, channel, and domain mismatch.
- Cross-genre training and testing Protocols (CGPs)
  - Genre groups. Ten genres are randomly grouped
  - CGPs for experiments
    - Seen genres: training & testing
    - Unseen: only testing

Group	Genre Types
Group I	drama (dr), vlog (vl), speech (sp)
Group II	entertainment (en), interview (in), play (pl)
Group III	live broadcast (lb), movie (mo)
Group IV	singing (si), recitation (re)

CGP	Seen Genres	Unseen Genres
CGP I	Group I, Group II, Group III	Group IV
CGP II	Group I, Group II, Group IV	Group III
CGP III	Group I, Group III, Group IV	Group II
CGP IV	Group II, Group III, Group IV	Group I

- **Preliminary experimental result of ECAPA-TDNN**

Protocol	Training dataset	CNCeleb.Eval		CNComplex	
		SV-EER	SASV-EER	SV-EER	SASV-EER
CGP I	CNCeleb 1&2	9.38	-	10.05	37.64
	Combination	9.02	-	9.56	36.97
CGP II	CNCeleb 1&2	10.04	-	10.85	40.76
	Combination	9.75	-	10.79	40.17
CGP III	CNCeleb 1&2	10.12	-	10.67	39.62
	Combination	9.33	-	10.22	38.49
CGP IV	CNCeleb 1&2	9.59	-	10.24	37.97
	Combination	9.21	-	9.86	37.42

ECAPA-TDNN struggles when facing spoofing, channel, and domain mismatches at the same time.

- **Result on channel mismatch scenario**

	dr	en	in	lb	re	si	sp	vl
dr	+1.65	-1.23	+1.72	+5.97	-	-	-	-
en	+2.71	+3.06	-0.02	-1.04	-7.35	+3.83	+1.20	-1.06
in	+4.32	+2.13	+2.49	-0.64	+2.70	+2.72	+3.94	+14.06
lb	-1.74	-1.13	-2.91	+0.02	-	+7.73	-	+6.71
sp	+2.98	+0.36	+3.34	+5.17	-	+2.07	-0.79	-
vl	-6.52	+0.51	+4.75	+2.63	-	-2.92	-	+2.01

Our approach achieves EER reductions compared to the PLDA baseline, demonstrating its effectiveness in addressing channel mismatch scenarios.

- **Result on spoofing attack scenario**

Protocol	Training dataset	CNCeleb.Eval		CNComplex	
		SV-EER	SASV-EER	SV-EER	SASV-EER
CGP I	CNCeleb 1&2	7.96	-	8.52	7.37
	Combination	7.79	-	7.56	7.25
CGP II	CNCeleb 1&2	8.24	-	8.85	8.57
	Combination	7.96	-	8.34	8.47
CGP III	CNCeleb 1&2	8.43	-	9.07	8.52
	Combination	8.19	-	8.82	8.25
CGP IV	CNCeleb 1&2	8.23	-	8.68	7.73
	Combination	8.13	-	8.45	7.48

This table shows our approach consistently outperforms ECAPA across all protocols, highlighting its robustness against spoofing attacks, channel mismatch, and domain mismatch.

- **Result on domain mismatch scenario**

Table 6. EER (%) of experimental results on CNCeleb.Eval testing dataset for the scenario of domain mismatch. For each protocol, the baseline system is established using the proposed model trained through the straightforward supervised learning paradigm. A bold number indicates the highest performance of this genre. The presence of an unseen group is indicated within brackets

Protocol	System	Overall	Group I			Group II			Group III		Group IV	
			dr	vl	sp	en	in	pl	lb	mo	si	re
CGP I (Group IV)	Baseline	21.31	23.15	20.77	15.86	21.61	22.99	27.35	17.50	29.32	28.63	14.58
	proposed approach	17.42	**22.66**	**16.44**	**14.46**	**20.95**	**19.47**	**20.93**	**16.31**	**24.57**	**22.33**	**13.89**
CGP II (Group III)	Baseline	22.10	22.66	20.64	16.41	23.56	24.08	27.29	19.43	31.55	25.25	**13.17**
	proposed approach	18.48	**18.04**	**18.08**	**13.84**	**19.11**	**20.10**	**19.71**	**17.84**	**27.16**	**24.16**	13.91
CGP III (Group II)	Baseline	23.46	25.42	23.89	17.58	26.38	25.34	29.25	21.72	31.14	27.74	**13.10**
	proposed approach	20.02	**23.27**	**20.70**	**16.79**	**22.08**	**22.83**	**22.96**	**19.54**	**28.53**	**22.92**	13.55
CGP IV (Group I)	Baseline	22.59	24.13	23.47	16.29	**19.50**	22.95	27.81	19.99	28.12	24.87	11.33
	proposed approach	19.98	**22.35**	**19.44**	**14.58**	21.34	**19.62**	**20.87**	**17.83**	**24.24**	**23.53**	**10.55**

Conclusions

- The proposed framework better addresses spoofing, channel, and domain mismatches than traditional ASV systems.
- We see the potential of an integrated, multi-task learning approach for real-world speaker verification.

[1] Zeng, Chang, et al. "Improving Generalization Ability of Countermeasures for New Mismatch Scenario by Combining Multiple Advanced Regularization Terms." Interspeech 2023

[2] Zeng, Chang, et al. "Attention Backend for Automatic Speaker Verification with Multiple Enrollments." ICASSP 2022