Spoofing-aware Speaker Verification System Robust Against Domain And Channel Mismatches

Chang Zeng

National Institute of Informatics & SOKENDAI

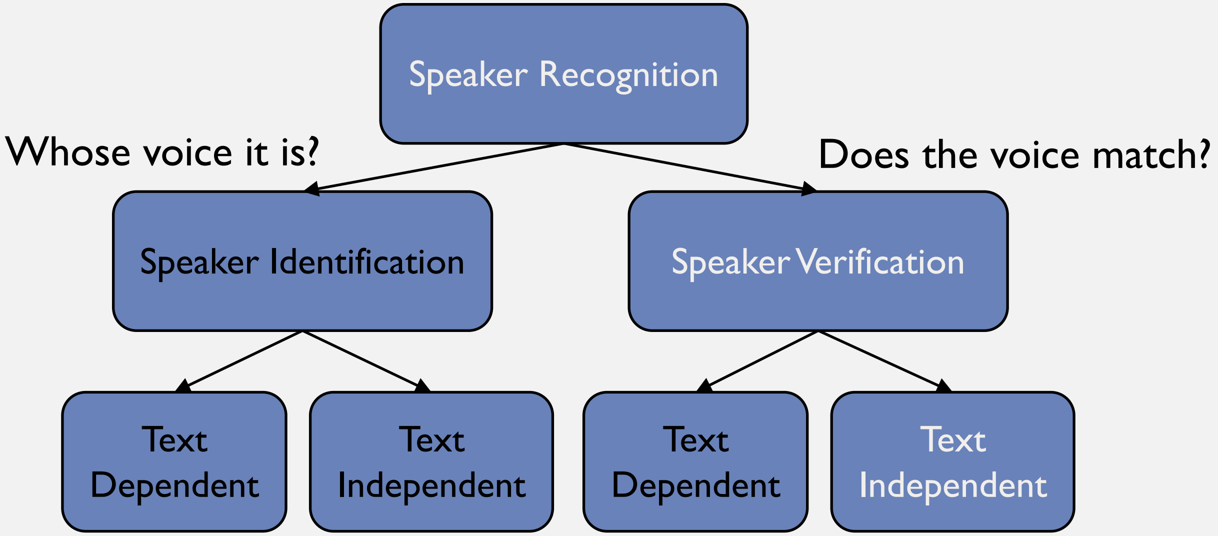
Yamagishi Lab

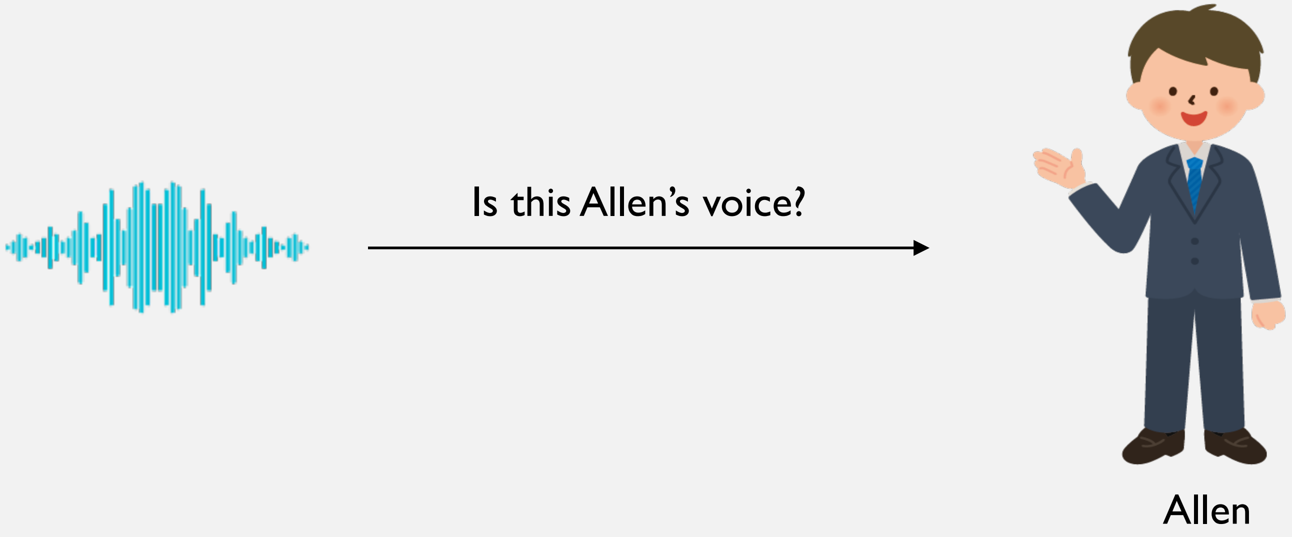
January 26, 2024

CONTENTS

INTRODUCTION

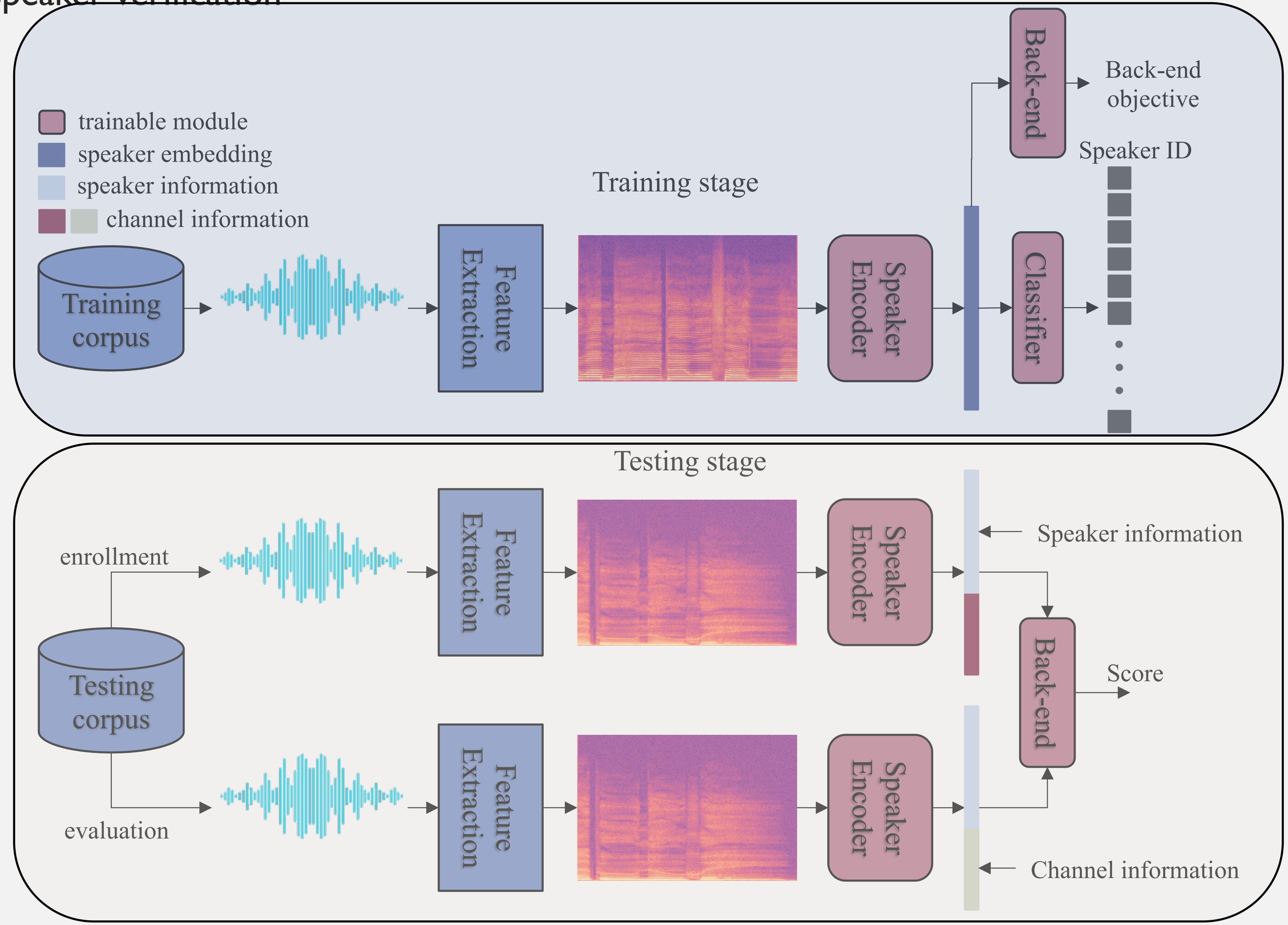
- What is "speaker verification"



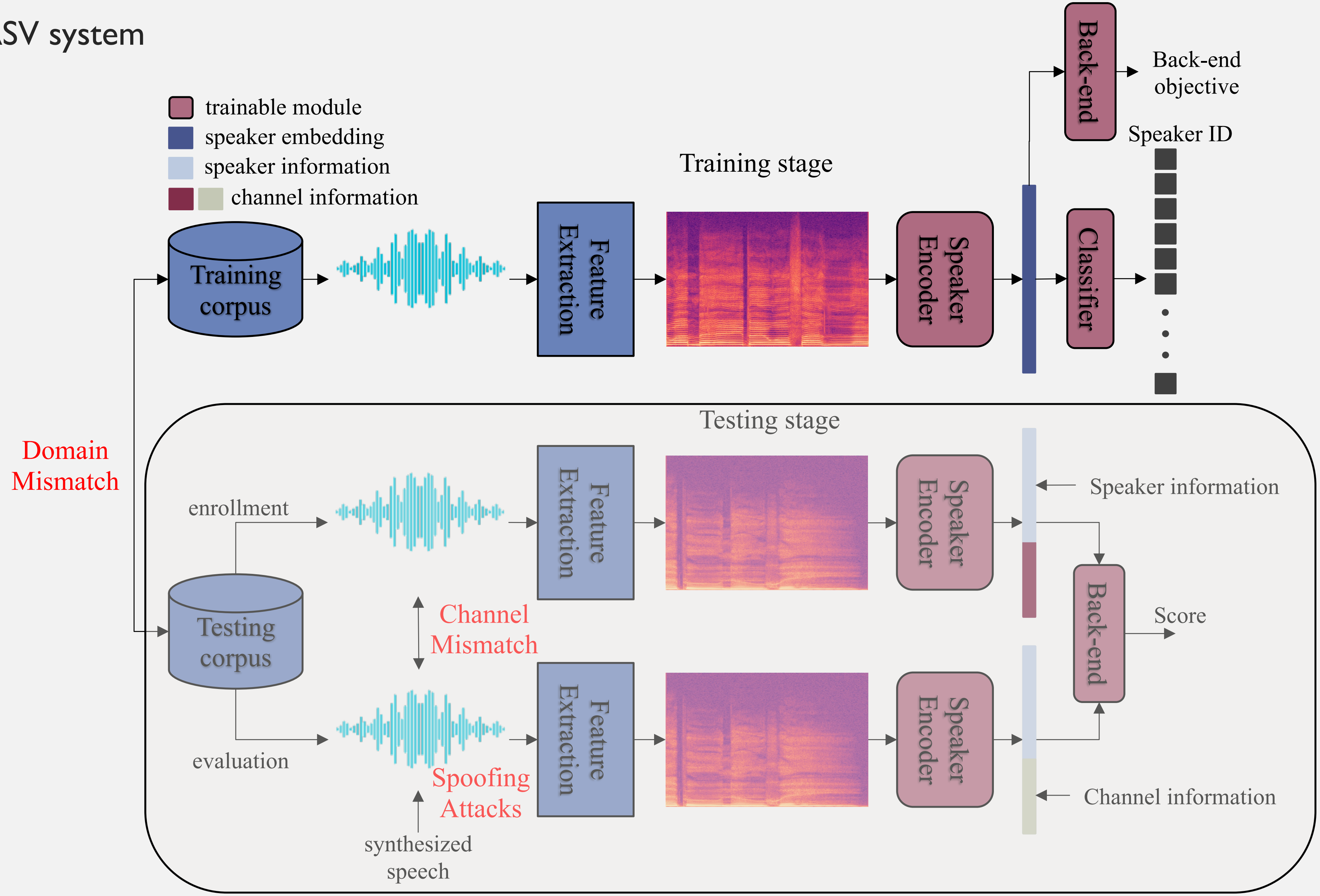


INTRODUCTION

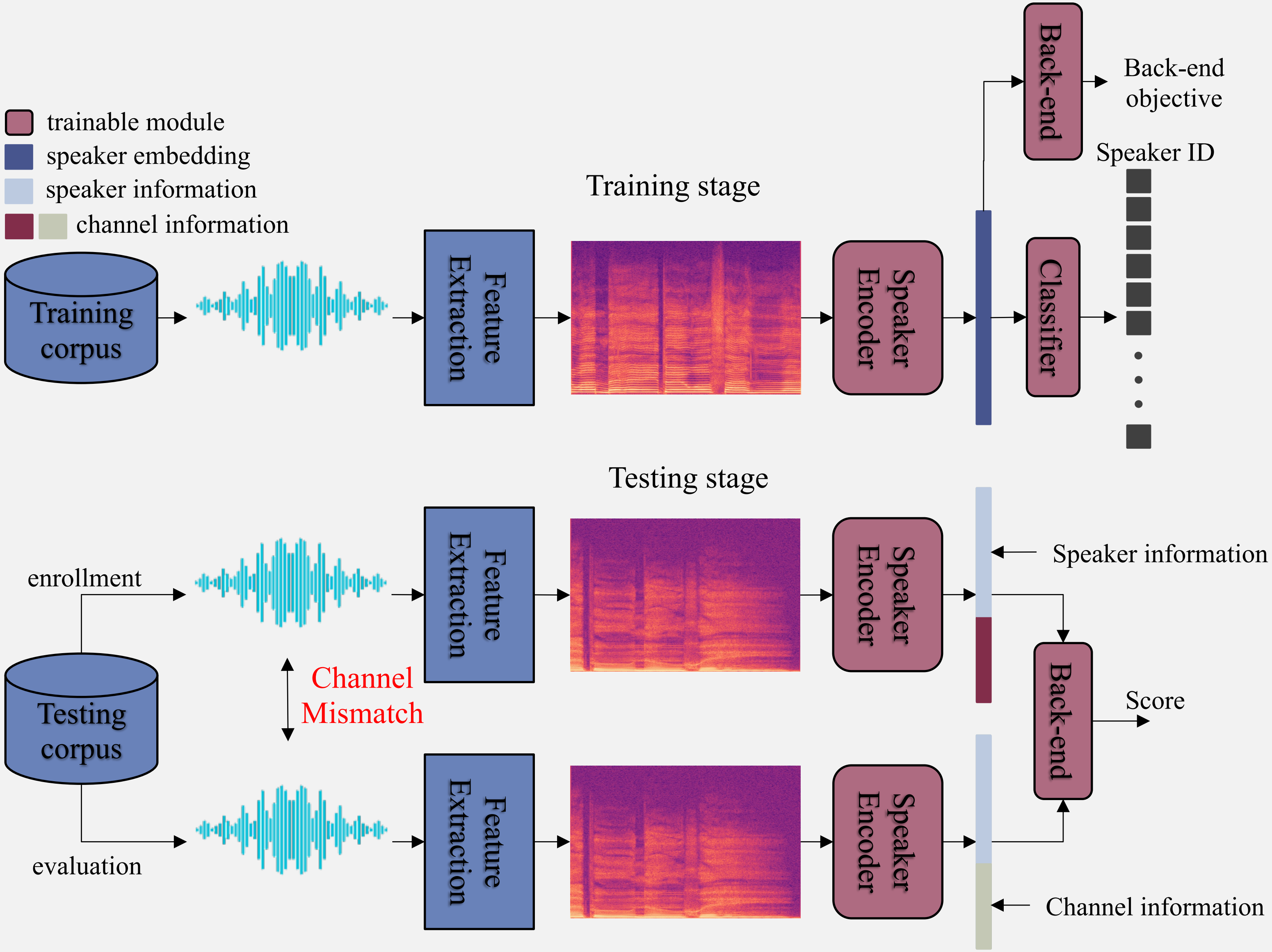
- Deep learning-based speaker verification



- Threats to ASV system



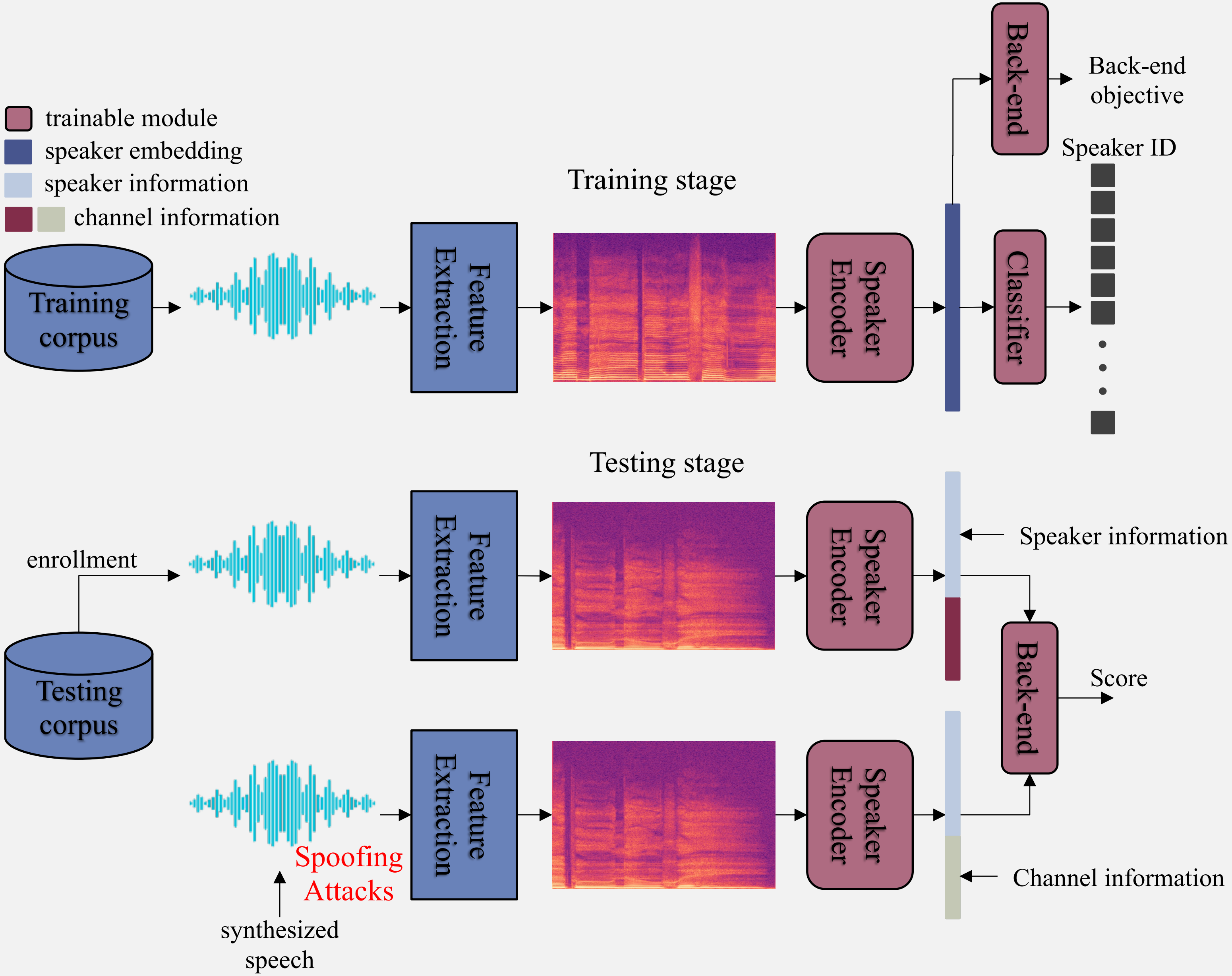
- Challenges of ASV



INTRODUCTION

- Issue 1: Channel Mismatch of ASV (Testing stage)
  - Channel variation leads to a mismatch between enrollment and evaluation utterances, referred to as channel mismatch, which is a prevalent issue for ASV systems. Note: both enrollment and evaluation utterances are provided by users.
- Examples of channel variation
  - Communication channel (microphone, telephone, …)
  - Acoustic environment (vlog, speech, interview, …)
  - Recording device (Android, iPhone, …)
  - Language (Chinese, Japanese, …)

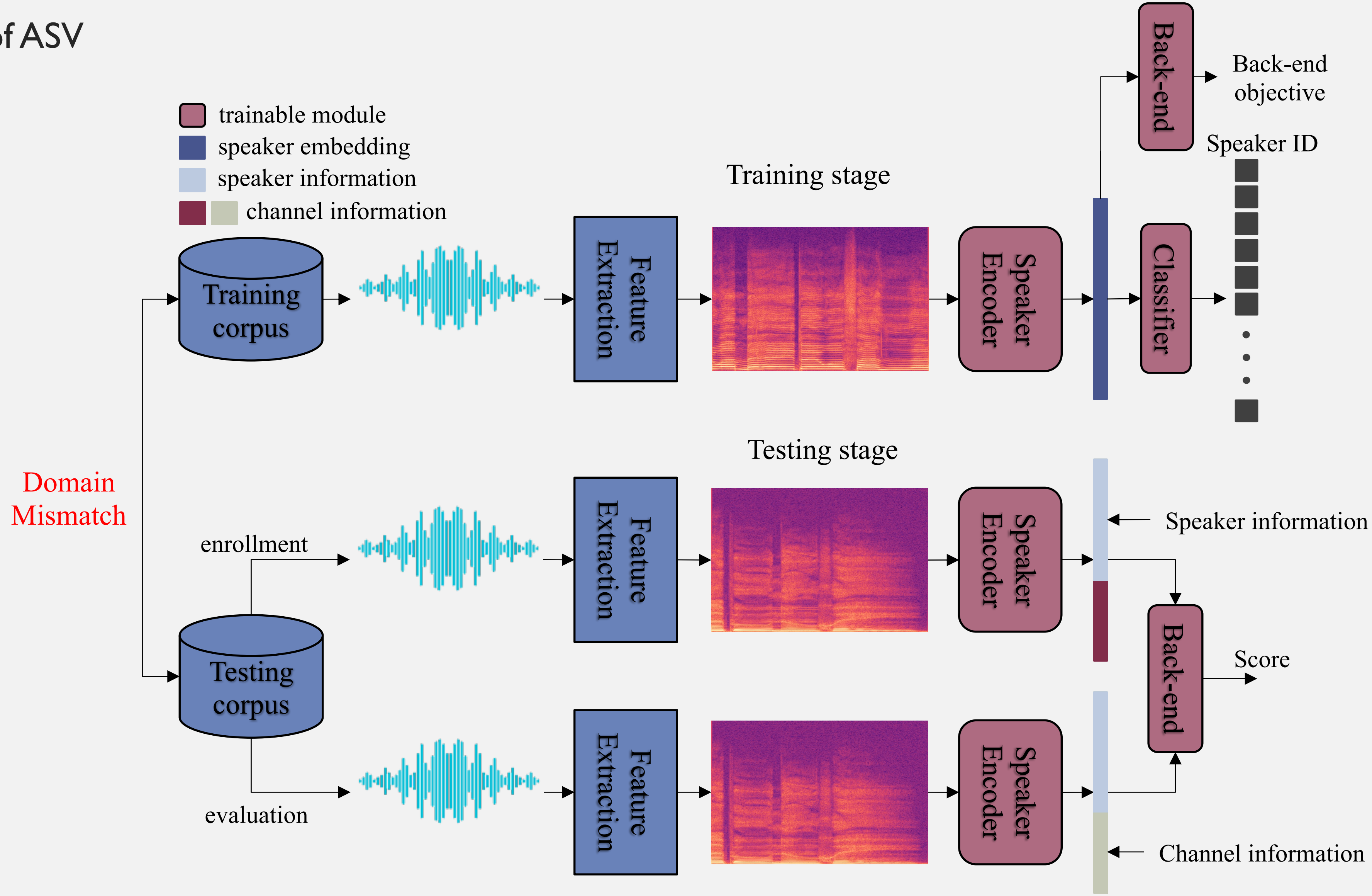
- Challenges of ASV



INTRODUCTION

- Issue 2: Spoofing Attacks to ASV (Testing stage)
  - Spoofing attacks generated by speech synthesis methods, including text-to-speech (TTS) and voice conversion make ASV system vulnerable.
- Examples of spoofing attacks
  - Multi-speaker TTS
    - Multi-speaker Tacotron
    - FastSpeech 1&2
    - VITS 1&2
    - VALL-E
  - Voice conversion
    - Voice clone
    - StyleGAN 1&2

- Challenges of ASV



INTRODUCTION

- Issue 3: Domain Mismatch of ASV (Training stage and testing stage)
  - Domain mismatch is a prevalent concern within the machine learning community, denoting a disparity between training and testing data distributions.
  - ASV systems also suffer from the challenges posed by domain mismatch.
- Examples of domain mismatch
  - Cross-age[1]
  - Cross-genre[2]
  - Cross-microphone[3]

1. Qin, X., Li, N., Chao, W., Su, D., Li, M. (2022) Cross-Age Speaker Verification: Learning Age-Invariant Speaker Embeddings. Proc. Interspeech 2022, 1436-1440, doi: 10.21437/Interspeech.2022-648
2. H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang and H. Chen, "Learning Domain-Invariant Transformation for Speaker Verification," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 7177-7181, doi: 10.1109/ICASSP43922.2022.9747514.
3. H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang and H. Chen, "Meta-Learning for Cross-Channel Speaker Verification," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 5839-5843, doi: 10.1109/ICASSP39728.2021.9413978.

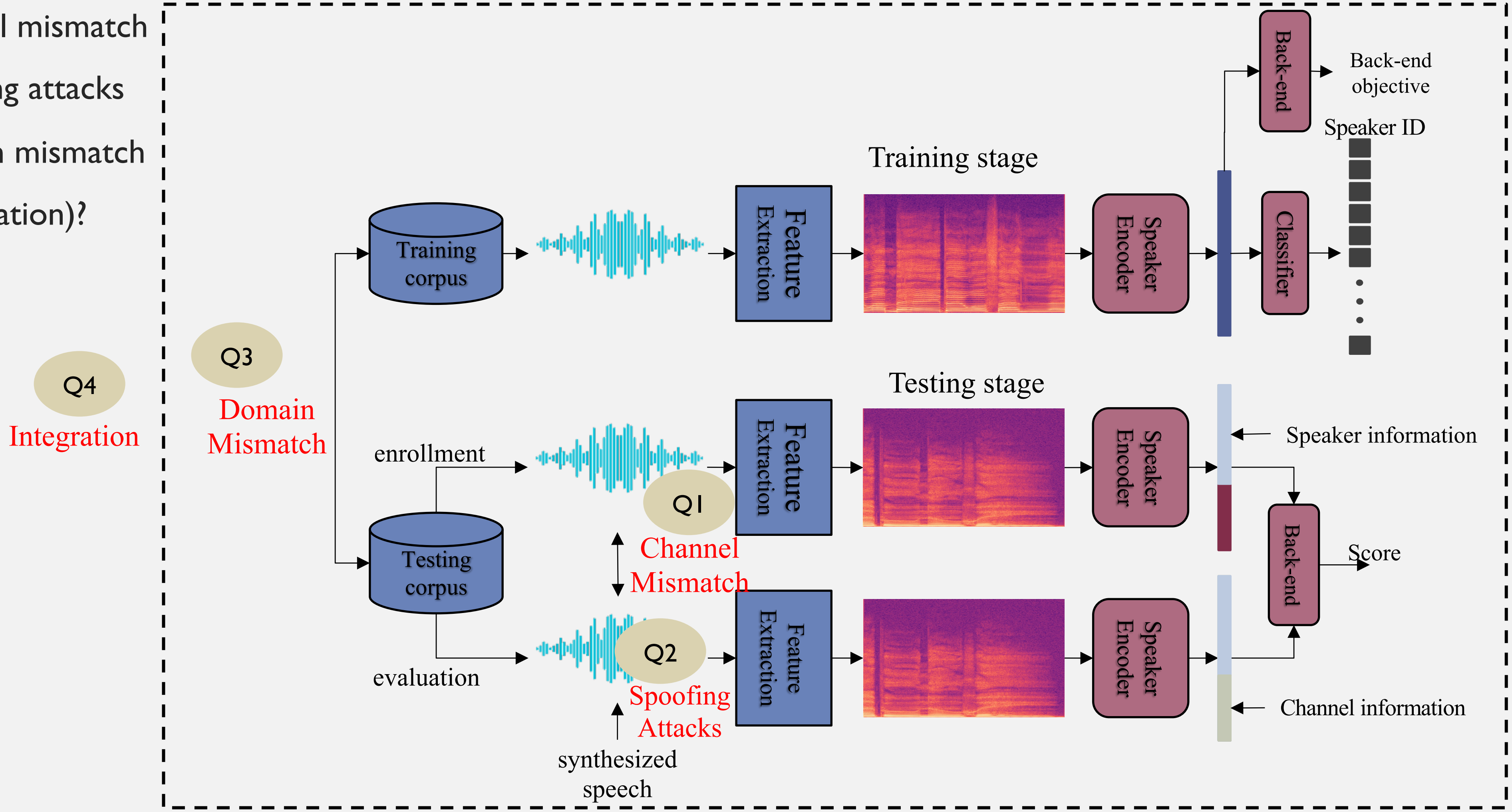
INTRODUCTION

- Summary: channel mismatch, spoofing attacks, domain mismatch
- These issues have been studied independently
  - Channel mismatch
    - NIST Speaker Recognition Challenge[1, 2] series have explored the channel mismatch of speech from the microphone and telephone.
  - Spoofing attacks
    - ASVspoof Challenge[5,6,7,8] series has explored the countermeasures (CMs) resisting the spoofing attacks for the ASV model.
    - Spoof-Aware Speaker Verification Challenge (SASVC)[9] attempts to integrate the CMs sub-model and ASV sub-model to make the ASV system spoof-aware.
  - Domain mismatch
    - Far Field Speaker Verification Challenge[3,4] series have explored the cross-microphone speech from the far and near field.

1. George R Doddington, Mark A Przybocki, Alvin F Martin, and Douglas A Reynolds. The nist speaker recognition evaluation–overview, methodology, systems, results, perspective. Speech communication, 31(2-3):225–254, 2000.
2. Seyed Omid Sadjadi, Craig S Greenberg, Elliot Singer, Douglas A Reynolds, Lisa P Mason, Jaime Hernandez Cordero, et al. The 2019 nist speaker recognition evaluation cts challenge. In Odyssey, pages 266–272, 2020.
3. Qin, X., Li, M., Bu, H., Rao, W., Das, R. K., Narayanan, S., & Li, H. (2020). The INTERSPEECH 2020 Far-Field Speaker Verification Challenge. Proc. Interspeech 2020, 3456–3460.
4. Qin, X., Li, M., Bu, H., Narayanan, S., & Li, H. (2022). Far-field Speaker Verification Challenge (FFSVC) 2022 : Challenge Evaluation Plan.
5. Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In Sixteenth annual conference of the international speech communication association, 2015.
6. Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. 2017
7. Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans,Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language, 64:101114, 2020.
8. Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. arXiv preprint arXiv:2109.00537, 2021
9. Jee weon Jung, Hemlata Tak, Hye jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. SASV 2022: The First Spoofing-Aware Speaker Verification Challenge. In Proc. Interspeech 2022, pages 2893–2897, 2022.

RESEARCH QUESTION

- Issue 4: Integration (robust against three threats simultaneously)
  - Can we build an ASV system
    - Q1: Robust against channel mismatch
    - Q2: Robust against spoofing attacks
    - Q3: Robust against domain mismatch
    - Q4: simultaneously (integration)?



RESEARCH QUESTION

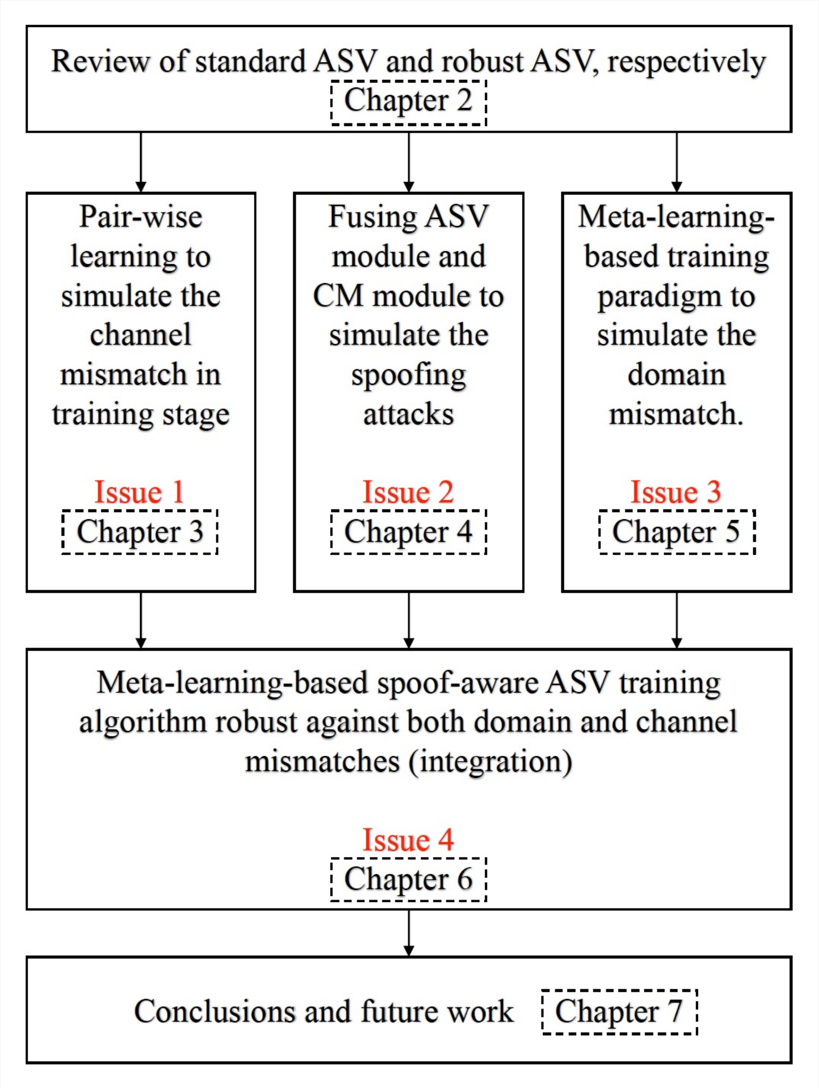
- Yes, we can build an ASV system robust against three threats simultaneously!
- Motivation and Importance
  - It is common that m ... is robust to only one threat may fail wher...
  - Building such a syste ... maintained in practical conditions with mult...

Domain mismatch scenario

Table 6.4: 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 means the best performance of this genre.

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
	Our 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**
	Our 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**
	Our 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
	Our approach	19.98	**22.35**	**19.44**	**14.58**	21.34	**19.62**	**20.87**	**17.83**	**24.24**	**23.53**	**10.55**

THESIS OUTLINE



THESIS SETTINGS

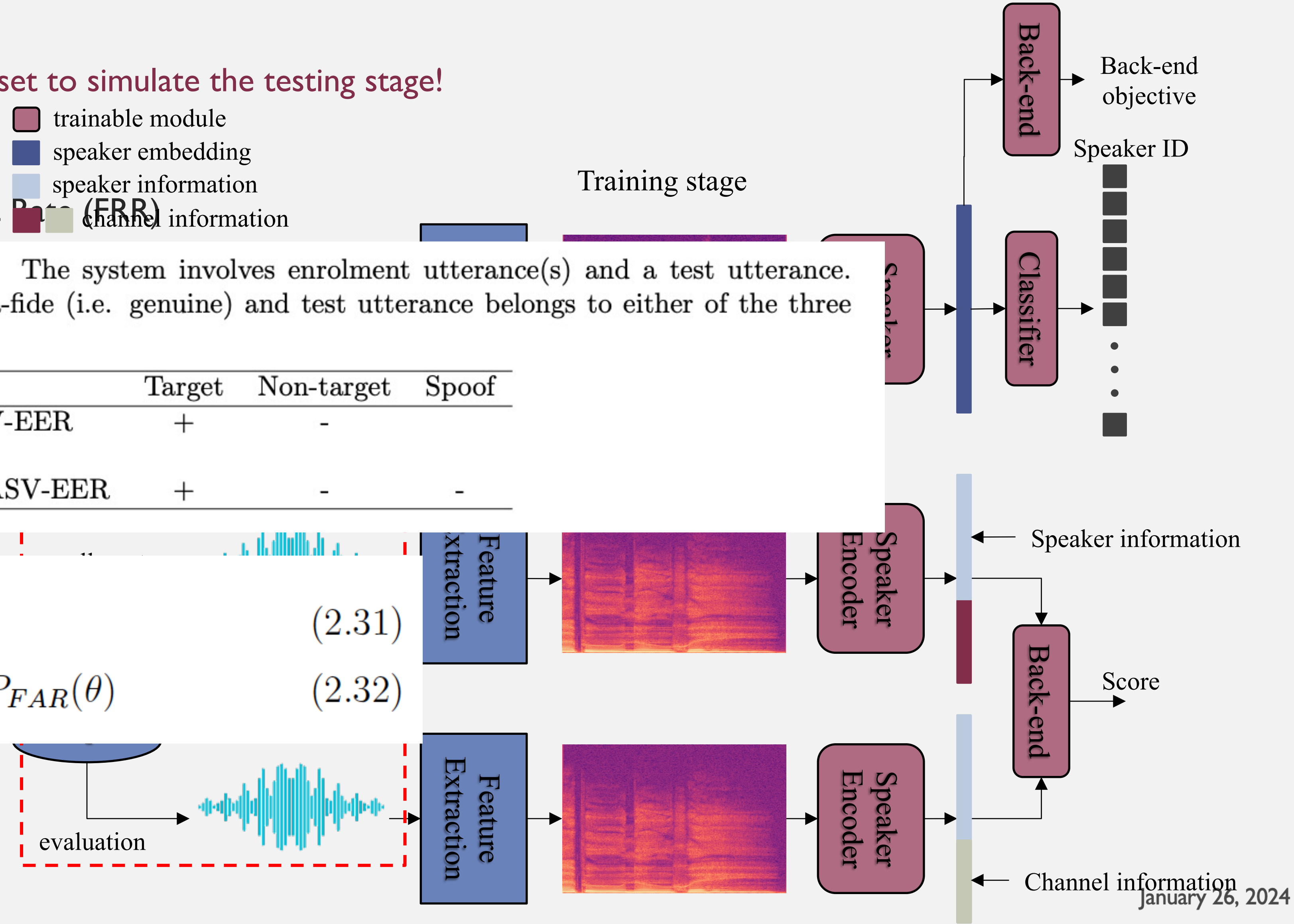
- Trial: the tuple of the evaluation utterance and the claimed enrolled speaker's identity constitutes a trial.
  - Testing trial
  - Training trial -> Sampling from training dataset to simulate the testing stage!
- Evaluation metrics
  - EER: False Accept Rate (FAR) = False Reject Rate (FRR)
    - SV-EER (sp
    - SASV-EER
  - minDCF
    - Cost coeff
    - Prior prob

Table 1: Description of EERs. The system involves enrolment utterance(s) and a test utterance. Enrolment utterance(s) is bona-fide (i.e. genuine) and test utterance belongs to either of the three types.

	Target	Non-target	Spoof
SV-EER	+	-	
SASV-EER	+	-	-

$$C_{det}(\theta) = C_{FR} * P_{target} * P_{FRR}(\theta) \quad (2.31)$$

$$+ C_{FA} * (1 - P_{target}) * P_{FAR}(\theta) \quad (2.32)$$

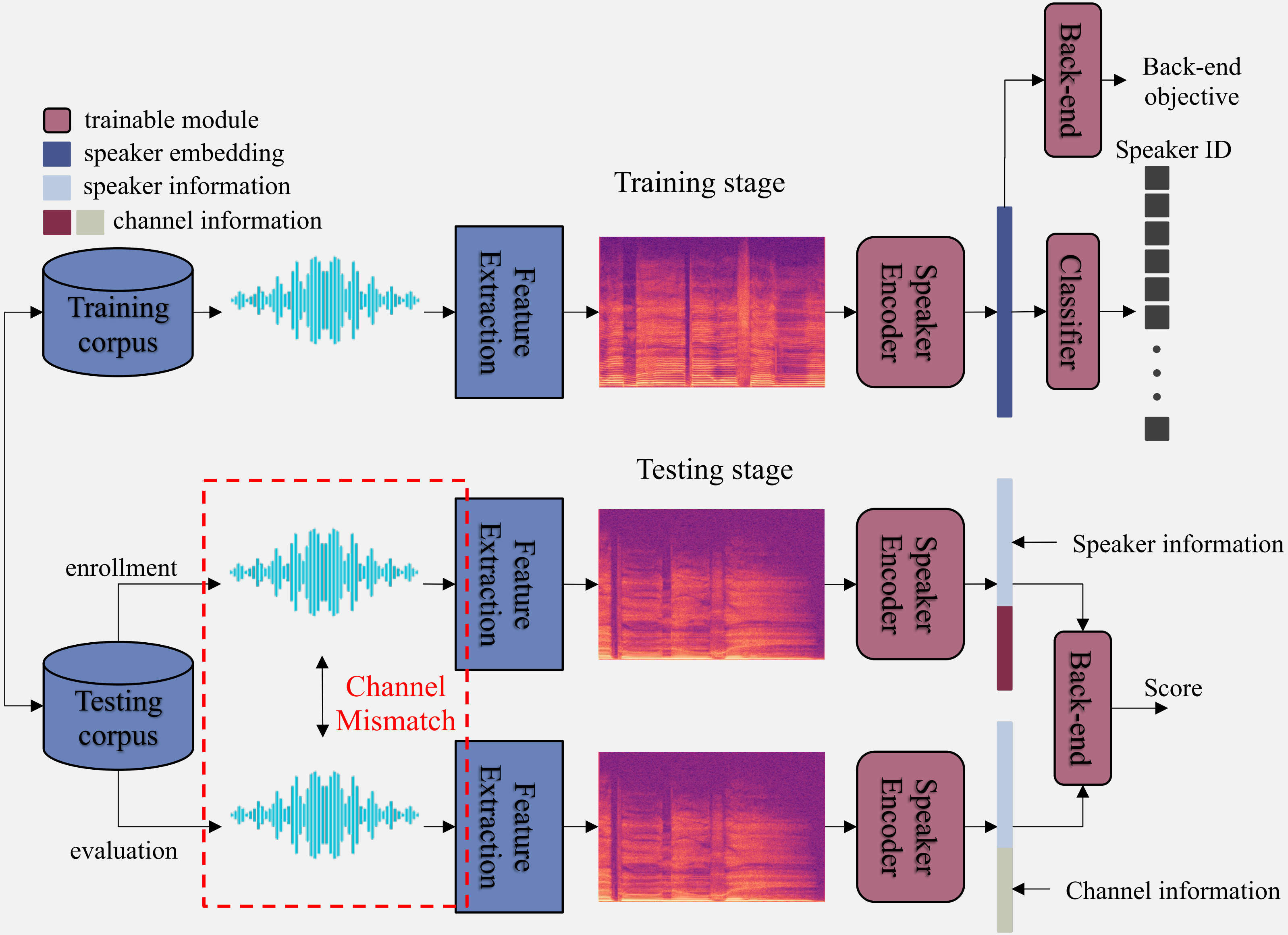


THESIS SETTINGS

- Study case in this thesis
  - Using the "genre" as an example to illustrate channel and domain mismatch
    - (Testing) Channel mismatch: The genres between enrollment and evaluation utterances are distinct.
    - (Training-Testing) Domain mismatch: The testing dataset has unseen genres that don't exist in the training dataset.
    - Difference: In terms of channel mismatch, all data are provided by users. However, for domain mismatch, training data are collected by the system creator, and testing data are provided by users.
  - Using the "copy-synthesis" method to create spoofing attacks via vocoders based on real Mel spectrogram
    - WORLD
    - Griffin-lim
    - Parallel WaveGAN
    - Multi-band MelGAN
    - HiFiGAN

CONTENTS

ISSUE 1: CHANNEL MISMATCH

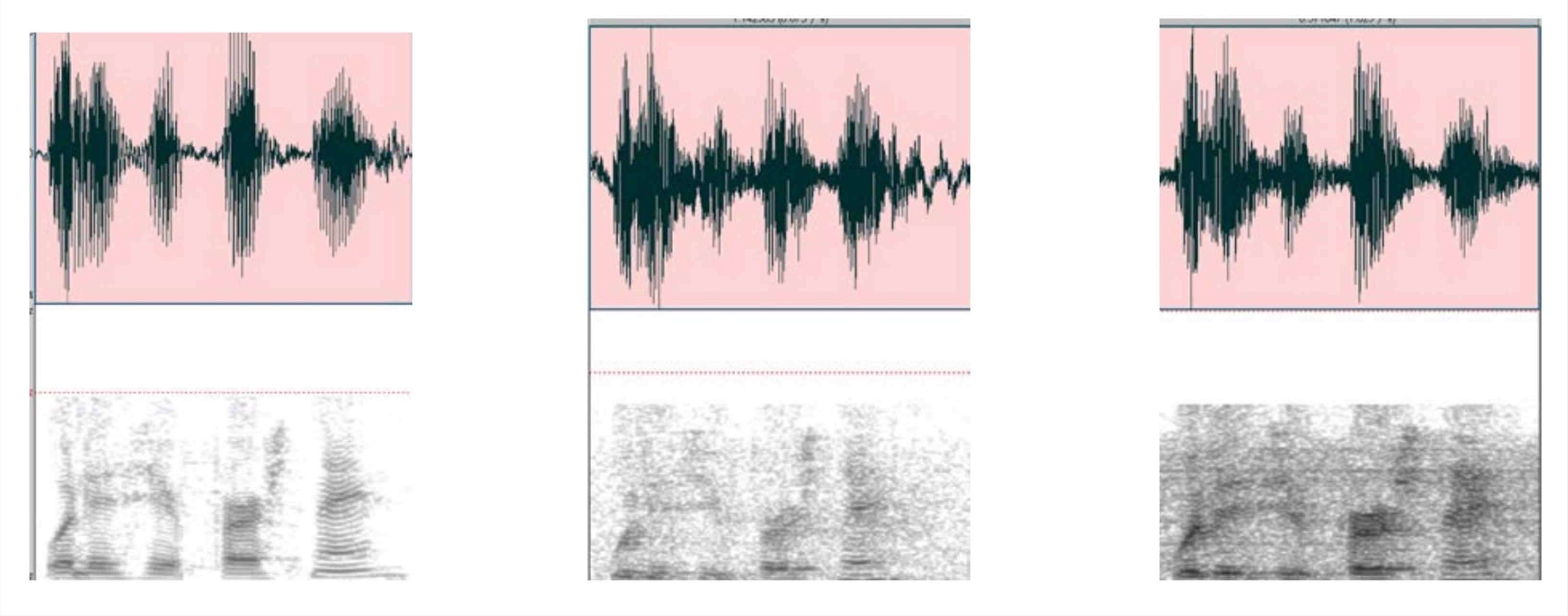


ISSUE I: CHANNEL MISMATCH

- Background
  - Channel variation leads to mismatch between enrollment and evaluation utterances

Speaker A + Channel Y ≠ Speaker A + Channel X

  - Example of the same signal recorded simultaneously with different devices



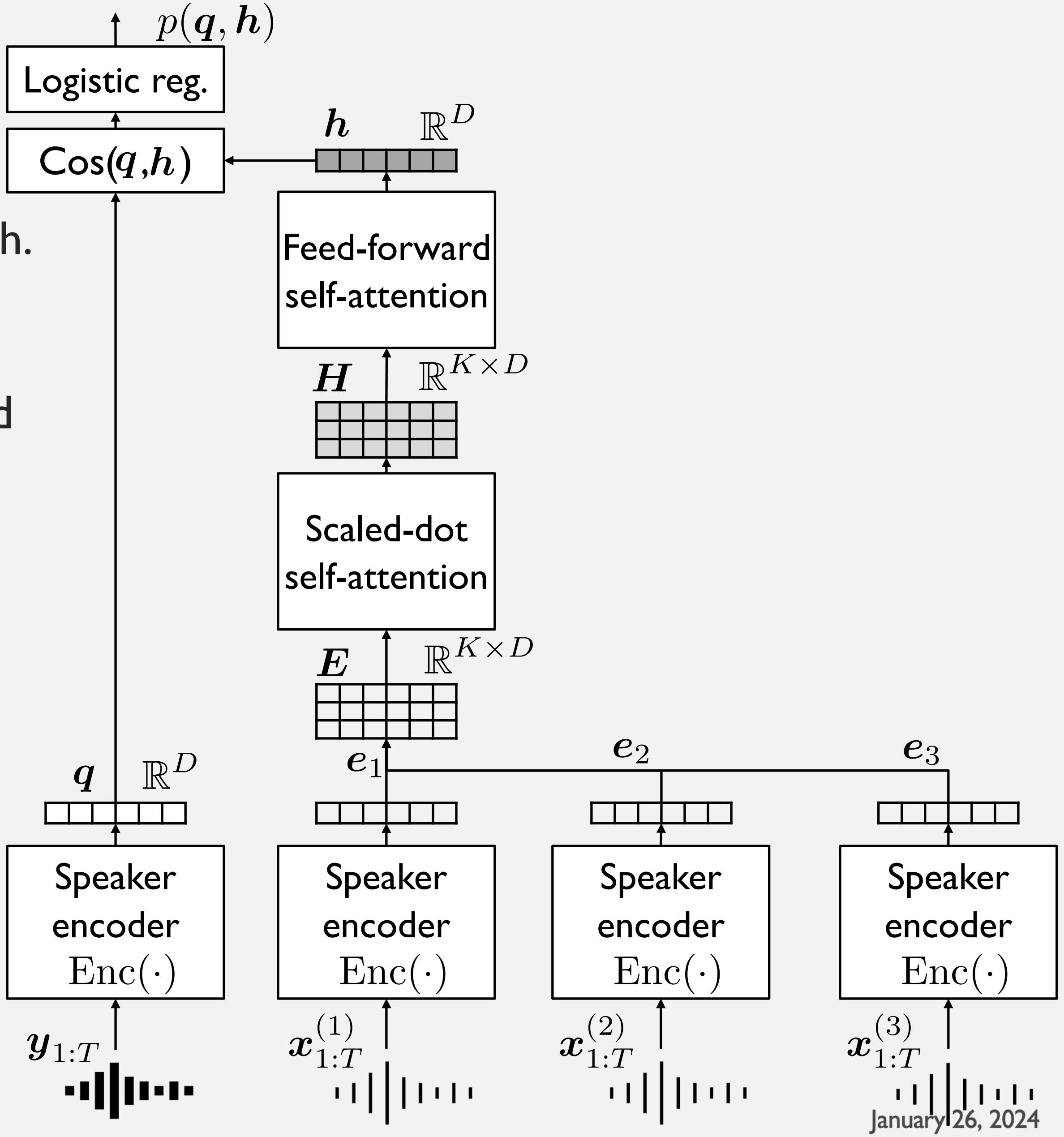
This figure comes from a lecture given by Prof. Kong Aik Lee.

ISSUE 1: CHANNEL MISMATCH

- Related work
  - Learning disentanglement representation[1,2]
    - Purifying the speaker embedding by removing channel information
    - Probabilistic Linear Discriminant Analysis (PLDA)
  - Pair-wise learning paradigm[3]
    - Simulating channel mismatch scenarios in the training stage for robustness.
    - Metric learning
  - Fusion of disentanglement representation and pair-wise learning[4]
    - NPLDA
- The limitation of the previous work
  - For multiple enrollment case, simply average or concatenate speaker embeddings of multiple enrollment utterances
  - There is no work to consider how to make good use of multiple enrollment utterances -> multiple enrollment utterances can cover more variations!

1. Ioffe, Sergey. "Probabilistic linear discriminant analysis." *European Conference on Computer Vision*. Springer, Berlin, Heidelberg, 2006.
2. Kenny, Patrick, and Pierre Dumouchel. "Disentangling speaker and channel effects in speaker verification." *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE, 2004.
3. Cumani, Sandro, et al. "Pairwise Discriminative Speaker Verification in the I-Vector Space." *IEEE Transactions on Audio, Speech, and Language Processing* 21.6 (2013): 1217-1227.
4. Ramoji, Shreyas, Prashant Krishnan, and Sriram Ganapathy. "NPLDA: A deep neural PLDA model for speaker verification." *arXiv preprint arXiv:2002.03562* (2020).

- Motivation:
  - Using multiple enrollments to cover more variations.
  - Pair-wise learning paradigm to simulate channel mismatch.
- Model architecture
  - Extract enrollment speaker embeddings $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_K\}$ and testing speaker embedding $\boldsymbol{q}$
  - Stacking all enrollment speaker embeddings as matrix $\boldsymbol{E}$.
  - Exploring intra-relationships among all enrollment speaker embeddings
    - Scaled-dot self-attention (SDSA)
  - Aggregating a varying number of enrollment speaker embeddings by adaptive weights
    - Feed-forward self-attention (FFSA)
  - Score calibration
    - Logistic regression (LR) for score calibration



APPROACH: ATTENTION BACK-END

- Trials sampling method for training
  - Why:
    - Simulate channel mismatch
    - Introduce multiple enrollment process in the training stage
  - How:
    - Load speaker-balanced min-batch from dataset
    - Rearrange the mini-batch to form trial pairs (*evaluation, enrollments*)
    - Positive pair: evaluation and enrollment data are from the same speaker, one evaluation utterance is selected from the speaker's data, and the rest are left for enrollment.
    - Negative pair: pairs marked by evaluation = ✓, enroll=(×, ×, ×) of other speakers included in a mini-batch.

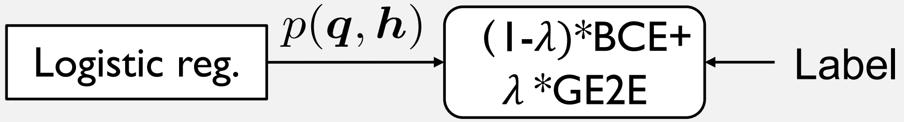
Table 1. Composition of pairs of (test-trial, enrollment-data) for training back-end model and ground-truth labels from mini-batch. A, B, and C are speaker IDs, and 1, 2, 3 and 4 are his or her audio IDs. ✓ and × denote test and enrollment audio files, respectively.

	A				*B*				*C*						
	1	2	3	4	1	2	3	4	1	2	3	4	Test	Enroll	Label
Trials to be used for training							⋮								

APPROACH: ATTENTION BACK-END

- Loss functions
  - A weighted sum of binary cross-entropy (BCE) loss and generalized end-to-end (GE2E) loss

$$\mathcal{L} = \lambda \mathcal{L}_{\text{ge2e}} + (1-\lambda)\mathcal{L}_{\text{bce}}$$



  - Binary cross-entropy loss

$$\mathcal{L}_{\text{bce}} = -\sum_{\forall l,m,n} [\mathcal{I}(l=n)\log P(\boldsymbol{q}_{lm}, \boldsymbol{h}_{nm}) + \mathcal{I}(l \neq n)\log(1-P(\boldsymbol{q}_{lm}, \boldsymbol{h}_{nm}))],$$

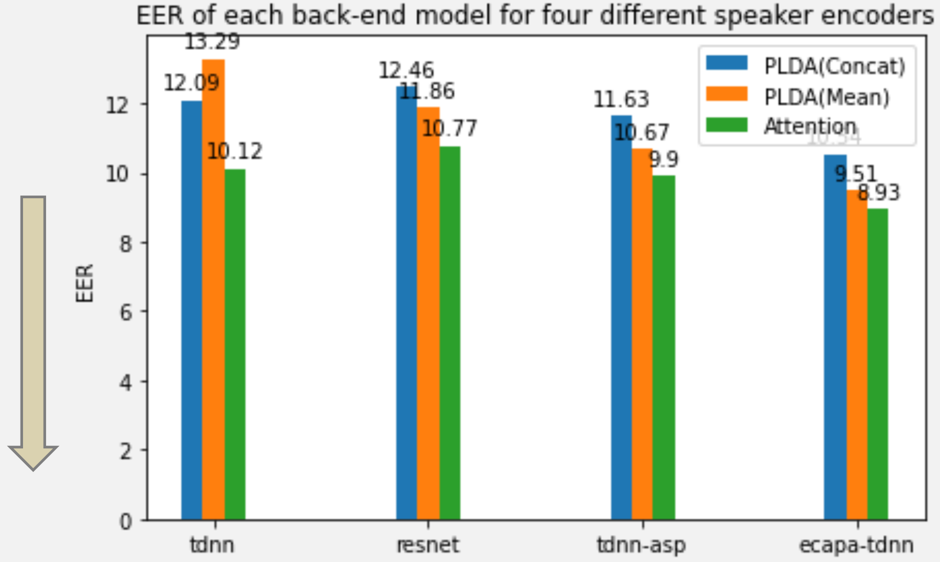
  - Generalized end-to-end loss

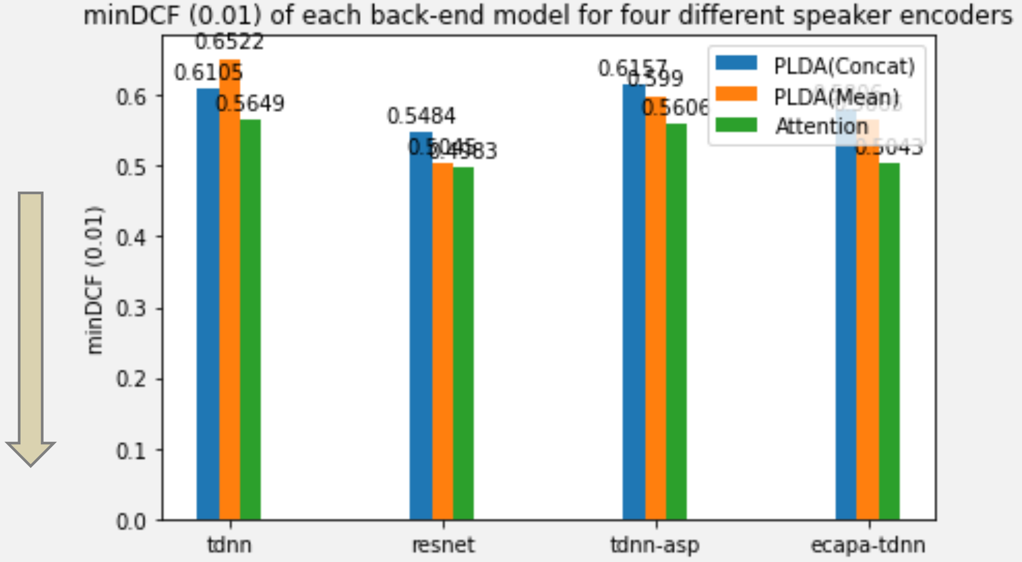
$$\mathcal{L}_{\text{ge2e}} = -\sum_{\forall l,m} \log \frac{\exp^{P(\boldsymbol{q}_{lm}, \boldsymbol{h}_{lm})}}{\sum_{\forall n} \exp^{P(\boldsymbol{q}_{lm}, \boldsymbol{h}_{nm})}}$$

Trials to be used for training	*A*				*B*				*C*						
	1	2	3	4	1	2	3	4	1	2	3	4	Test	Enroll	Label
	✓	×	×	×									$\boldsymbol{q}_{A1}$	$\boldsymbol{h}_{A1}$	P
	✓					×	×	×					$\boldsymbol{q}_{A1}$	$\boldsymbol{h}_{B1}$	N
	✓									×	×	×	$\boldsymbol{q}_{A1}$	$\boldsymbol{h}_{C1}$	N
	×	✓	×	×									$\boldsymbol{q}_{A2}$	$\boldsymbol{h}_{A2}$	P
		✓			×		×	×					$\boldsymbol{q}_{A2}$	$\boldsymbol{h}_{B2}$	N
		✓							×		×	×	$\boldsymbol{q}_{A2}$	$\boldsymbol{h}_{C2}$	N
							⋮								
	×	×	×									✓	$\boldsymbol{q}_{C4}$	$\boldsymbol{h}_{A4}$	N
					×	×	×					✓	$\boldsymbol{q}_{C4}$	$\boldsymbol{h}_{B4}$	N
									×	×	×	✓	$\boldsymbol{q}_{C4}$	$\boldsymbol{h}_{C4}$	P

APPROACH: ATTENTION BACK-END

- CNCeleb dataset
  - Contains 11 different genres of utterances, 2800 speakers
  - Official protocol has multiple enrollments
  - The number of enrollments is varying
- Experimental result
  - For Concat or Mean operation, we give the performance of PLDA model.
  - Despite the criterion, our proposed attention back-end realizes the best performance.





APPROACH: ATTENTION BACK-END

- Evidence for robustness to channel mismatch
  - New evaluation protocol with channel mismatch case based on CNCeleb test data;
  - Comparing the resistance to channel mismatch of PLDA and attention back-end

dr: drama
en: entertainment
in: interview
lb: live broadcast
re: recitation
si: singing
sp: speech
vl: vlog

Table 3.5: The difference of EER. The first row of the table denotes genres of testing utterances, and the first column of the table represents genres of enrollment utterances. The blue number means the attention back-end is better than PLDA. The red number means the attention back-end is worse than PLDA.

Evaluation genres (columns); Enrollment genres (rows)

	dr	en	in	lb	re	si	sp	vl
dr	+1.86	-2.24	+2.14	+3.93	-	-	-	-
en	+1.55	+1.68	+0.77	-1.65	-7.59	+4.39	+2.22	-2.49
in	+3.71	+1.42	+2.53	+0.51	+2.48	+1.42	+3.23	+12.97
lb	-2.39	+1.47	-2.59	+1.22	-	+7.66	-	+6.31
sp	+5.33	-1.05	+4.39	+6.39	-	+1.45	-1.22	-
vl	-8.33	+1.13	+6.38	+3.52	-	-2.86	-	+1.56

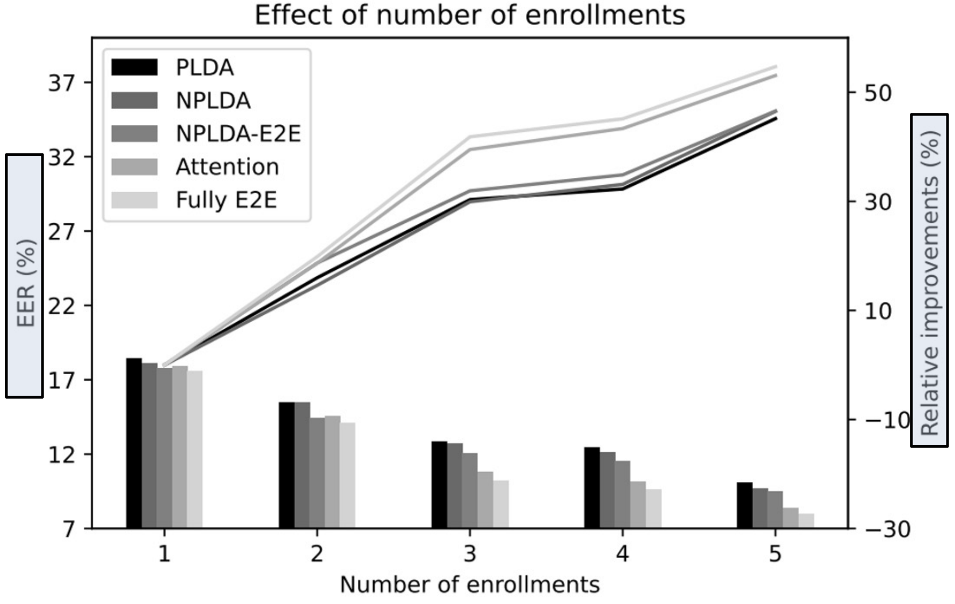
- Conclusion: Attention back-end is more robust than PLDA against channel mismatch!

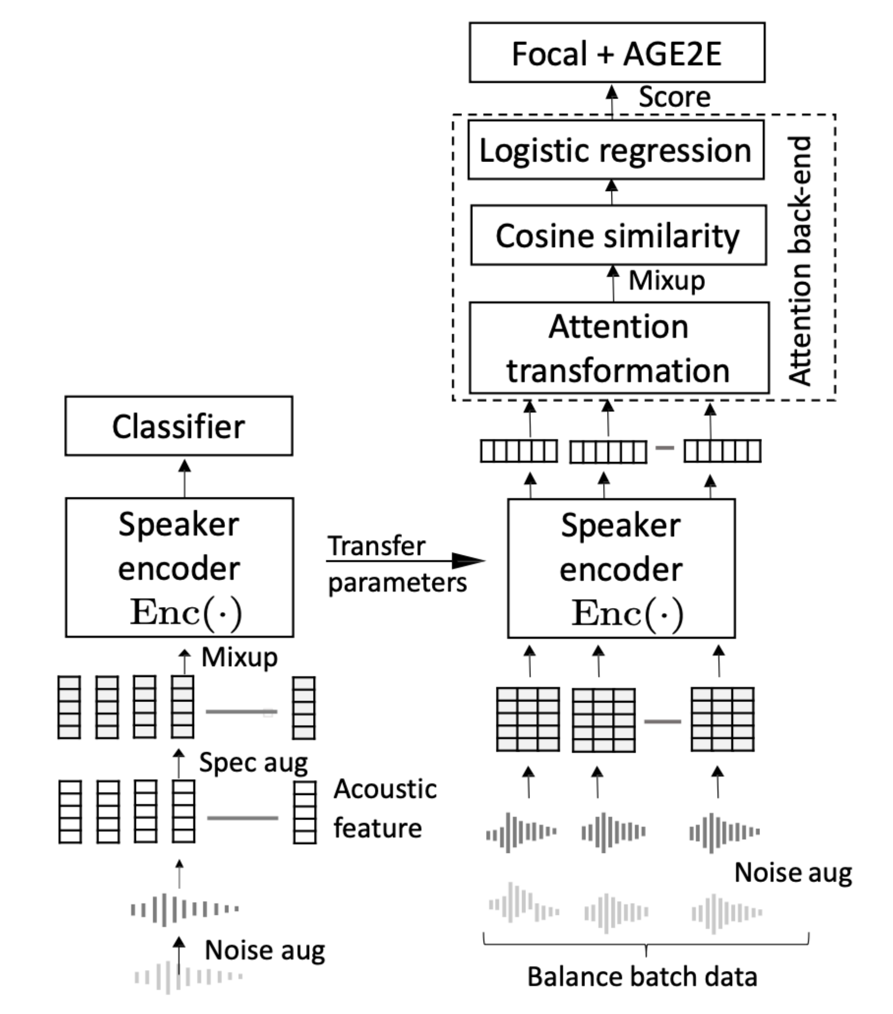
APPROACH: ATTENTION BACK-END

- More results from my journal paper (Fully E2E ASV model)[1]

The effect of the number of enrollment utterances

System	$K=1$	$K=2$	$K=3$	$K=4$	$K\geq 5$
ECAPA-TDNN/PLDA	18.44	15.49	12.85	12.49	10.11
ECAPA-TDNN/NPLDA	18.15	15.50	12.72	12.14	9.71
ECAPA-TDNN/NPLDA-E2E	17.76	14.44	12.09	11.57	9.49
ECAPA-TDNN/Attention backend	17.90	14.56	10.83	10.14	8.40
Fully E2E	**17.57**	**14.08**	**10.22**	**9.64**	**7.96**





1. **Chang Zeng**, Xiaoxiao Miao, Xin Wang, Erica Cooper, and Junichi Yamagishi. 2022. Joint Speaker Encoder and Neural Back-end Model for Fully End-to-End Automatic Speaker Verification with Multiple Enrollment Utterances. *arXiv preprint arXiv:2209.00485 (2022). Submitted to Computer Speech & Language.*

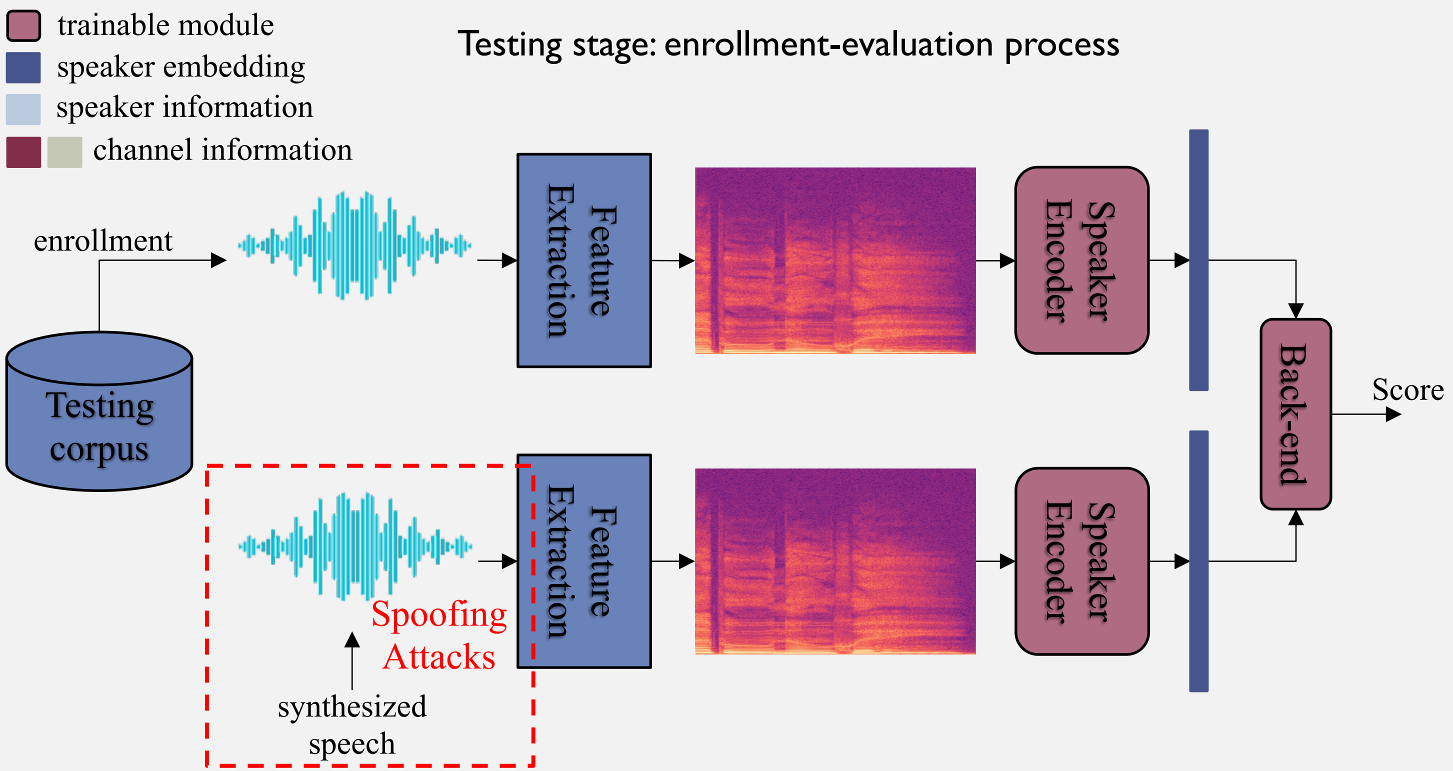
APPROACH: ATTENTION BACK-END

- Summary
  - The pair-wise learning paradigm can mitigate channel mismatch;
  - Attention back-end can better fuse multiple speaker embeddings with different channel information by attention mechanism to improve the channel robustness.

CONTENTS

ISSUE 2: SPOOFING ATTACKS

- Spoofing attacks



ISSUE 2: SPOOFING ATTACKS

- Background
  - Traditional ASV system is vulnerable when considering the scenario with spoofing attacks.

Model (EER)	SV-EER wo/ spoof attacks		SASV-EER w/ spoof attacks	
	Dev	Eval	Dev	Eval
ECAPA-TDNN[2]	1.88	1.63	17.38	23.83

  - Speech synthesis is abused by criminals.


Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find

Thomas Brewster Forbes Staff

Associate editor at Forbes, covering cybercrime, privacy, security and surveillance.

Follow

Oct 14, 2021, 07:01am EDT


Forbes

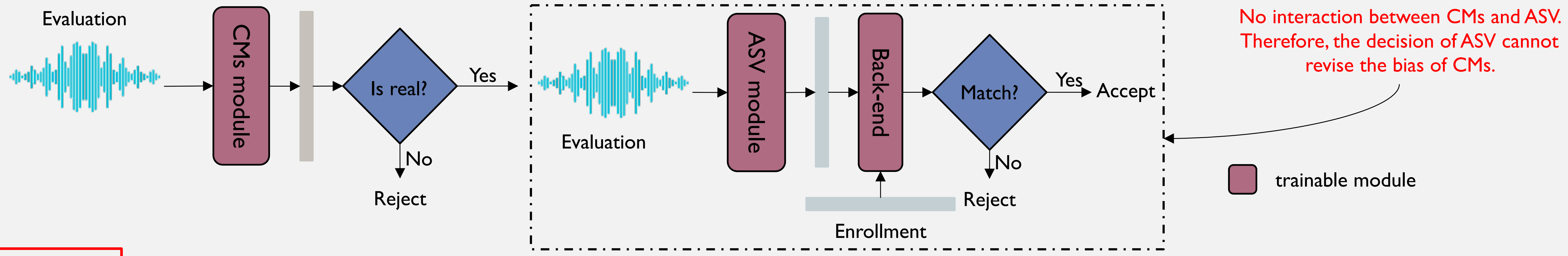
CONSUMER TECH

A Voice Deepfake Was Used To Scam A CEO Out Of $243,000

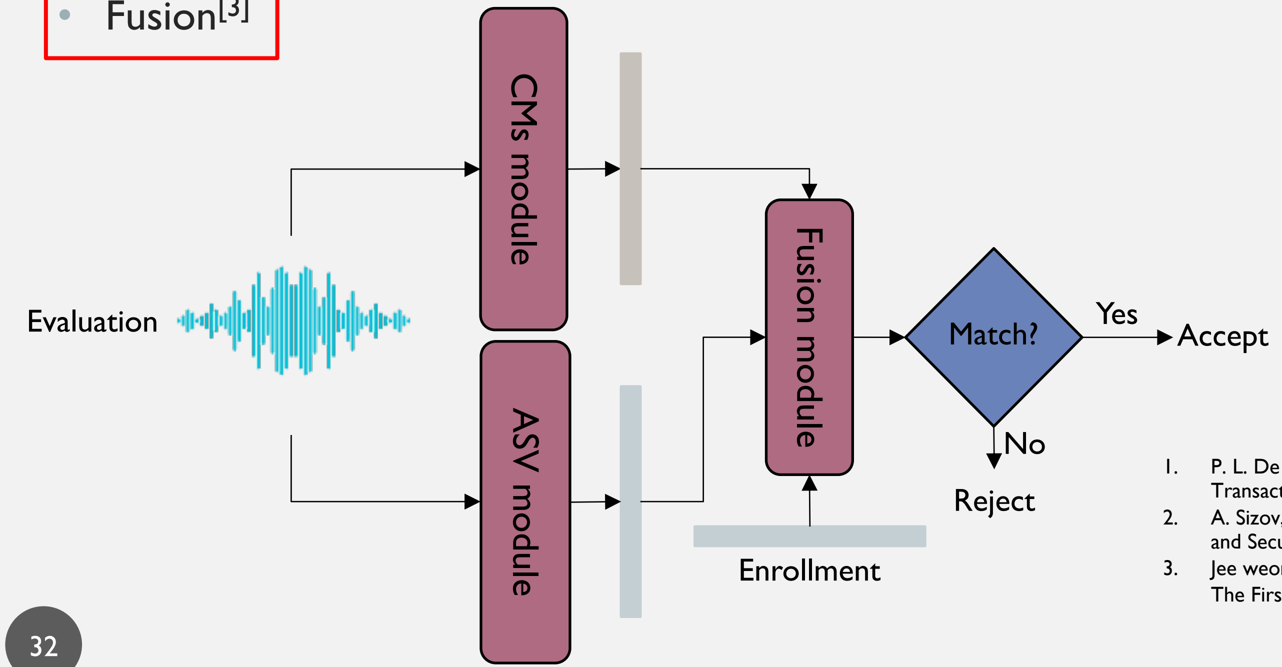
1. Jee weon Jung, Hemlata Tak, Hye jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. SASV 2022: The First Spoofing-Aware Speaker Verification Challenge. In Proc. Interspeech 2022, pages 2893–2897, 2022.
2. Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In Proc. Interspeech 2020, pages 3830–3834, 2020.

ISSUE 2: SPOOFING ATTACKS

- Related work
  - Cascade[1,2]: Independently training CMs and ASV module for two-stage decision



- Fusion[3]



1. P. L. De Leon, M. Pucher, J. Yamagishi et al., "Evaluation of speaker verification security and detection of hmm-based synthetic speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2280–2290, 2012.
2. A. Sizov, E. Khoury, T. Kinnunen et al., "Joint speaker verification and antispoofing in the i-vector space," IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, 2015.
3. Jee weon Jung, Hemlata Tak, Hye jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. SASV 2022: The First Spoofing-Aware Speaker Verification Challenge. In Proc. Interspeech 2022, pages 2893–2897, 2022.

APPROACH: SPOOFING-AWARE ATTENTION BACK-END

- Motivation
  - Improving robustness to both zero-effort impostor access attempts and spoofing attacks.
- Evaluation protocol for spoof-aware ASV scenario[1]
  - Positive case: test utterance matches enrollment utterance and test utterance is real
  - Negative cases:
    - Non-target (zero-effort imposter): the speaker of evaluation utterance is different from the speaker of enrollment utterance
    - Spoofing attacks: evaluation utterance is spoofed regardless of the similarity

Table 1: Description of EERs. The system involves enrolment utterance(s) and a test utterance. Enrolment utterance(s) is bona-fide (i.e. genuine) and test utterance belongs to either of the three types.

	Target	Non-target	Spoof
SV-EER	+	-	
SASV-EER	+	-	-

- Spoof-aware attention back-end[1]
  - Model architecture
    - Attention back-end
    - Calculate CM score P_{cm} and ASV score P_{asv}
  - Fusion module

$$P(P_{cm}, P_{asv}) = \frac{1}{1+\exp^{-s}} = \frac{1}{1+\exp^{-(w_1 * P_{cm} + w_2 * P_{asv} + v)}},$$

  - Loss function
    - Binary cross-entropy with N hardest negative samples

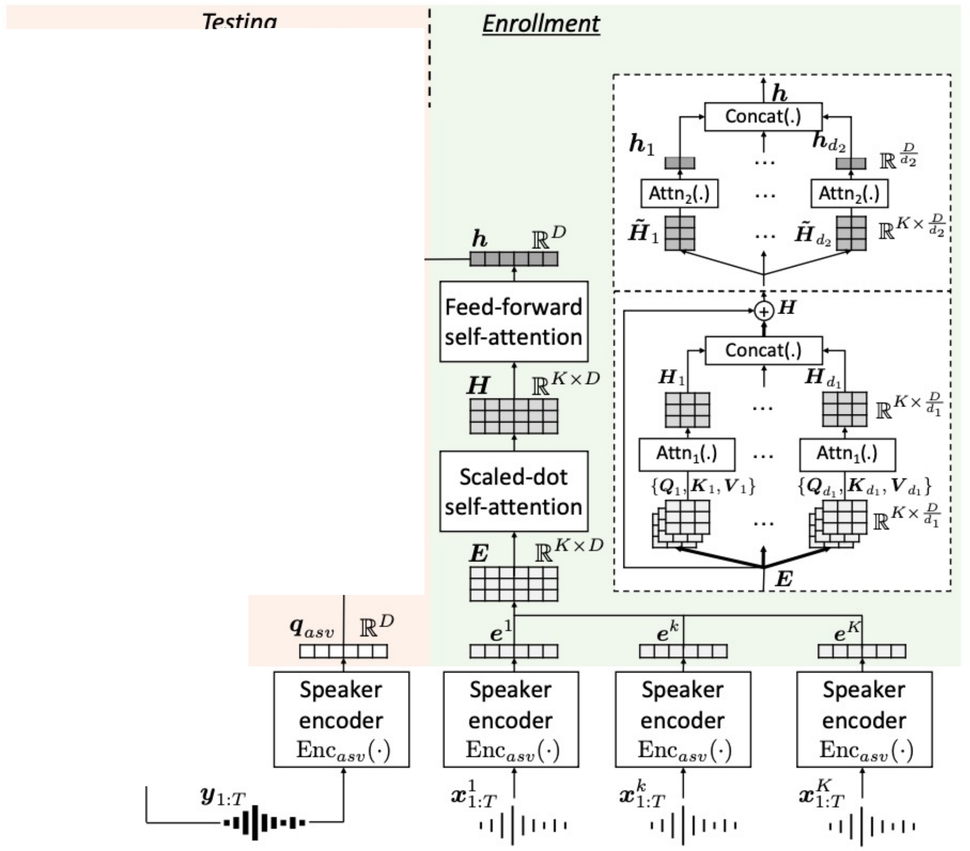


Figure 1: *Architecture of the proposed spoofing-aware attention back-end including score-level fusion.*

1. **Chang Zeng**, Lin Zhang, Meng Liu, and Junichi Yamagishi. Spoofing-Aware Attention based ASV Back-end with Multiple Enrollment Utterances and a Sampling Strategy for the SASV Challenge 2022. *In Proc. Interspeech 2022*, pages 2883–2887, 2022.

- ASVspoof19 dataset
  - Unbalanced real and spoof data
  - Provide ASV protocol with spoof evaluation data
  - Multiple enrollments
- Experimental result

Model (EER)	SV-EER wo/ spoof attacks		SASV-EER w/ spoof attacks	
	Dev	Eval	Dev	Eval
Attention back-end	1.54	1.42	16.78	22.91
Spoof-aware version	1.41	1.32	0.81	1.19

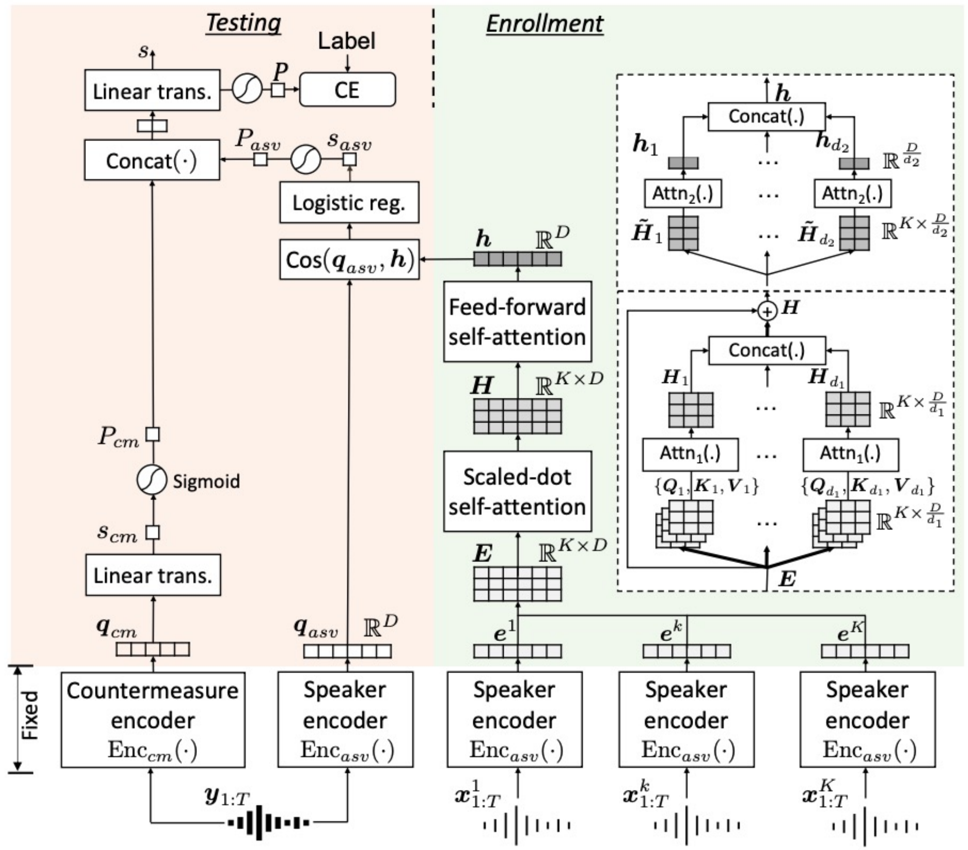


Figure 1: *Architecture of the proposed spoofing-aware attention back-end including score-level fusion.*

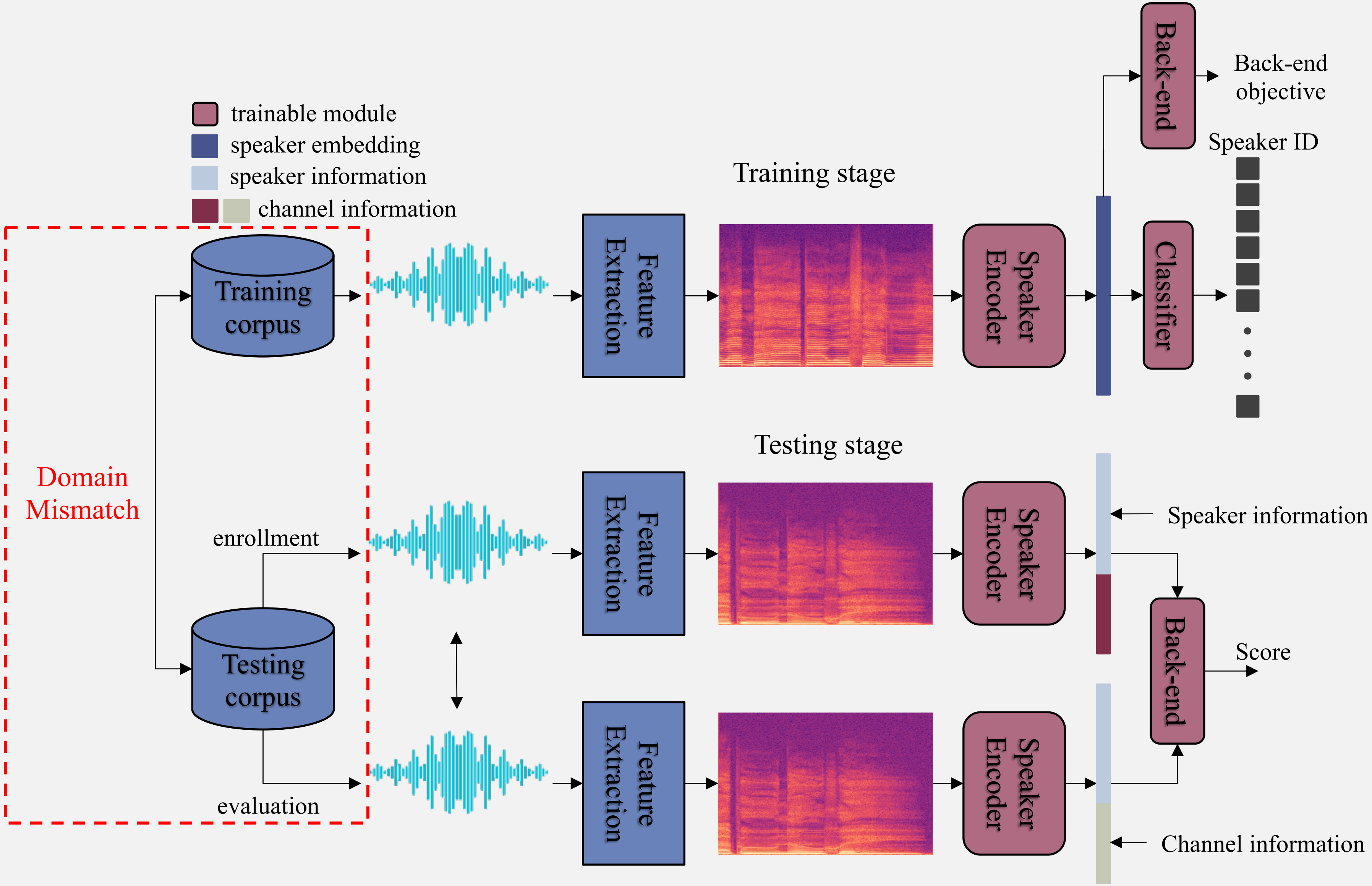
- Conclusion
  - After integrating CM information, attention back-end is much more robust than the one without CM information in the spoof-aware ASV scenario.

APPROACH: SPOOFING-AWARE ATTENTION BACK-END

- Summary
  - Attention back-end is a neural network, easily integrated with other modules.
  - Spoofing-aware attention back-end is an approach to issue1&2 by simulating both spoofing attacks and channel mismatch scenarios in the training stage

CONTENTS

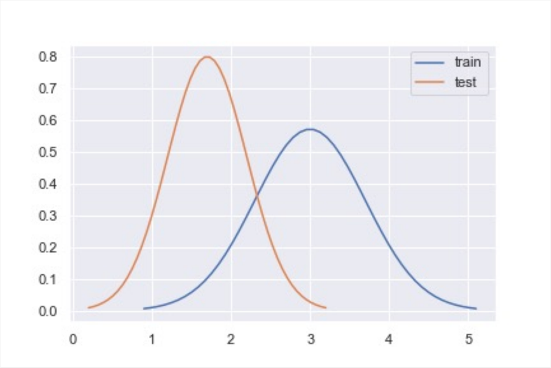
ISSUE 3: DOMAIN MISMATCH



ISSUE 3: DOMAIN MISMATCH

- Background of domain mismatch
  - Domain distribution shift is a general problem in machine learning
    - Independently identically distributed assumption
    - P(source domain or seen domain) ≠ P(target domain or unseen domain)

      training data — testing data



- As one of the most efficient, convenient, natural, and non-intrusive biometric characteristics, reliability is crucial and must be maintained in the face of domain mismatch scenario for the ASV system[1].

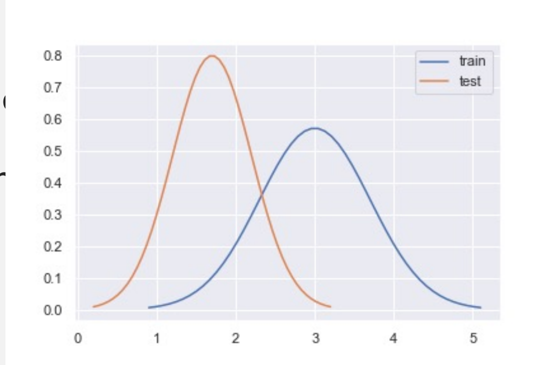
SV-EER

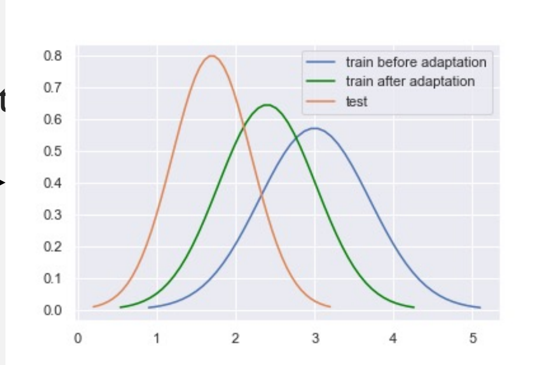
System	Spk encoder	Back-end	CNCeleb.E
TDNN[2]	VoxCeleb[3,4]	VoxCeleb	16.59
	VoxCeleb	CNCeleb	13.44
	CNCeleb[5,6]	CNCeleb	12.52

Domain mismatch!

1. Jee weon Jung, Hemlata Tak, Hye jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. SASV 2022: The First Spoofing-Aware Speaker Verification Challenge. In Proc. Interspeech 2022, pages 2893–2897, 2022.
2. Snyder, David, et al. "X-vectors: Robust dnn embeddings for speaker recognition." *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.
3. Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.
4. Chung, J. S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.
5. Fan, Yue, et al. "Cn-celeb: a challenging chinese speaker recognition dataset." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.
6. Li, L., Liu, R., Kang, J., Fan, Y., Cui, H., Cai, Y., ... & Wang, D. (2022). CN-Celeb: multi-genre speaker recognition. *Speech Communication*, *137*, 77-91.

- Related work of domain mismatch (in ASV)
  - Data augmentation: music, babble, noise, reverberation
  - Domain adaptation[1,2]
    - Supervised or unsupervised learning on in-… on to target
    - Methods: Adversarial-training-based, Recor… ased





    - Target-domain data is required
  - Domain generalization[3,4,5]
    - Learning closer distribution to target domain directly from training data, without further adaptation on target-domain data.
    - Methods: Gradient-based data manipulation, Representation disentanglement, Meta-learning paradigm
    - No target-domain data is required!

1. K. A. Lee, Q. Wang and T. Koshinaka, "The CORAL+ Algorithm for Unsupervised Domain Adaptation of PLDA," in Proc. ICASSP, 2019, pp. 5821-5825.
2. Q. Wang, K. Okabe, K. A. Lee and T. Koshinaka, "A Generalized Framework for Domain Adaptation of PLDA in Speaker Recognition," in Proc. ICASSP, 2020, pp. 6619-6623.
3. H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang and H. Chen, "Meta-Learning for Cross-Channel Speaker Verification," in Proc. ICASSP, 2021, pp. 5839-5843.
4. Kang, Jiawen, et al. "Domain-invariant speaker vector projection by model-agnostic meta-learning." *arXiv preprint arXiv:2005.11900* (2020).
5. Zhang, Hanyi, et al. "Learning Domain-Invariant Transformation for Speaker Verification." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

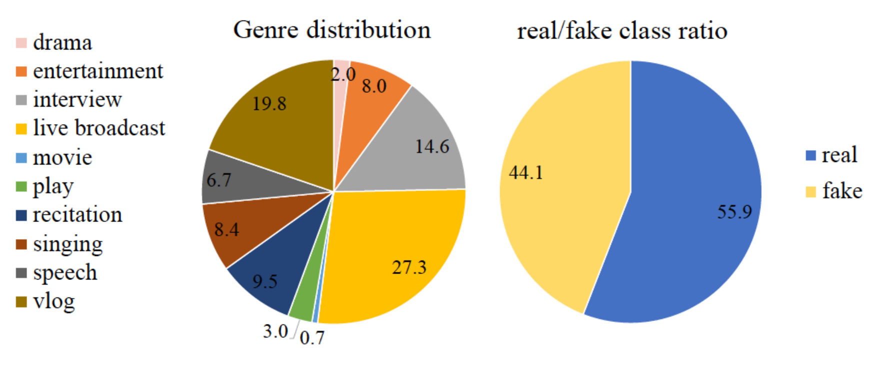
TASK

- Anti-spoofing
  - The successful application of the meta-learning algorithm in ASV tasks is evidenced by prior works[1,2,3].
  - Given the widespread nature of domain mismatch across various machine learning tasks, the method's effectiveness for fake audio detection is anticipated to extend successfully to the ASV task.

1. H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang and H. Chen, "Meta-Learning for Cross-Channel Speaker Verification," in Proc. ICASSP, 2021, pp. 5839-5843.
2. Kang, Jiawen, et al. "Domain-invariant speaker vector projection by model-agnostic meta-learning." *arXiv preprint arXiv:2005.11900* (2020).
3. Zhang, Hanyi, et al. "Learning Domain-Invariant Transformation for Speaker Verification." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

DATA PREPARATION

- CNSpoof dataset: Using the "copy-synthesis" method to create spoofing attacks via vocoders based on real Mel spectrogram
  - Griffin-Lim[1]
  - WORLD[2]
  - Parallel WaveGAN[3]
  - Multi-band MelGAN[4]
  - HiFi-GAN[5]



1. D. Griffin and Jae Lim, "Signal estimation from modified short-time Fourier transform," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236-243, April 1984.
2. Morise, Masanori, Fumiya Yokomori, and Kenji Ozawa. "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications." *IEICE TRANSACTIONS on Information and Systems* 99.7 (2016): 1877-1884
3. Yamamoto, Ryuichi, Eunwoo Song, and Jae-Min Kim. "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020
4. Yang, Geng, et al. "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech." *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021.
5. Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis." *Advances in Neural Information Processing Systems* 33 (2020): 17022-17033.

- Cross-genre protocols (CGP)
  - CNCeleb 1&2 + CNSpoof
  - Training data: only contains seen genres
  - Testing data: contains seen and unseen genres

Table 5.1: Genre group division

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)

Table 5.2: Cross-genre protocols (CGP)

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

- LCNN with simple supervised learning based on CGP I

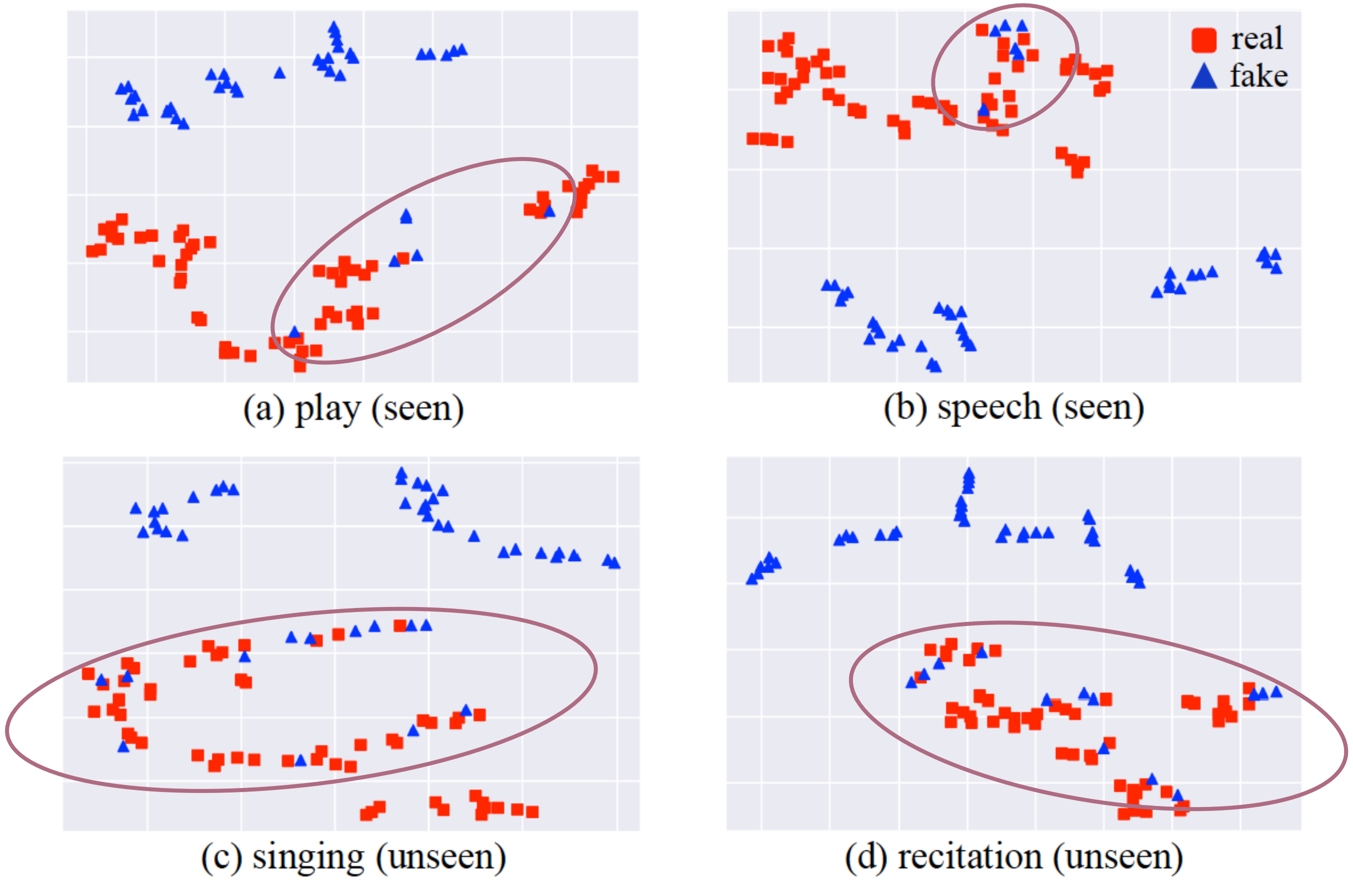
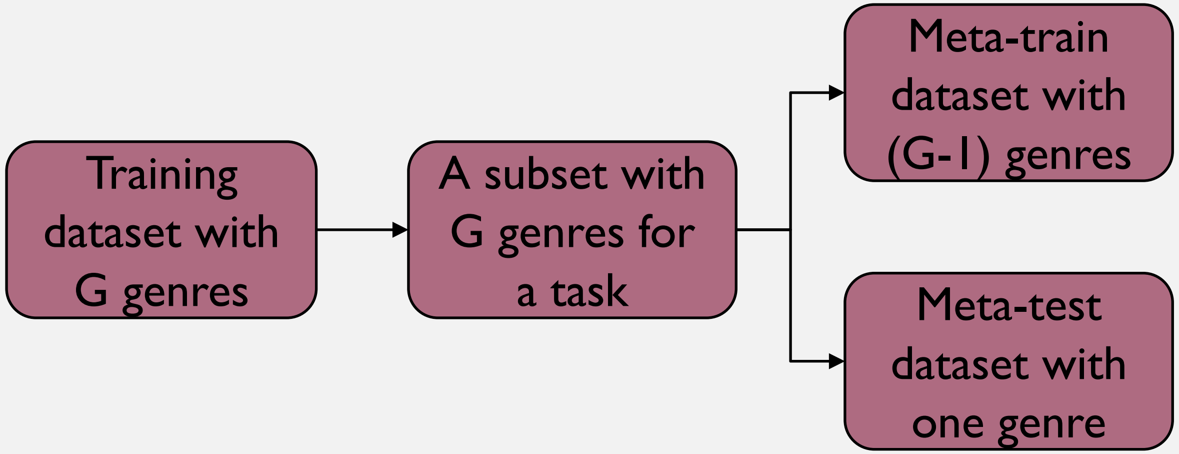


Figure 5.2: Visualization of LCNN countermeasure embedding by T-SNE.

META-LEARNING PARADIGM FOR ANTI-SPOOFING TASK

- Domain (Genre) sampling to construct a task set, which comprises T tasks
  - These tasks are created by randomly selecting data from the training dataset with G domains (genres)
  - Each task consists of two essential components
    - Meta-train dataset
    - Meta-test dataset



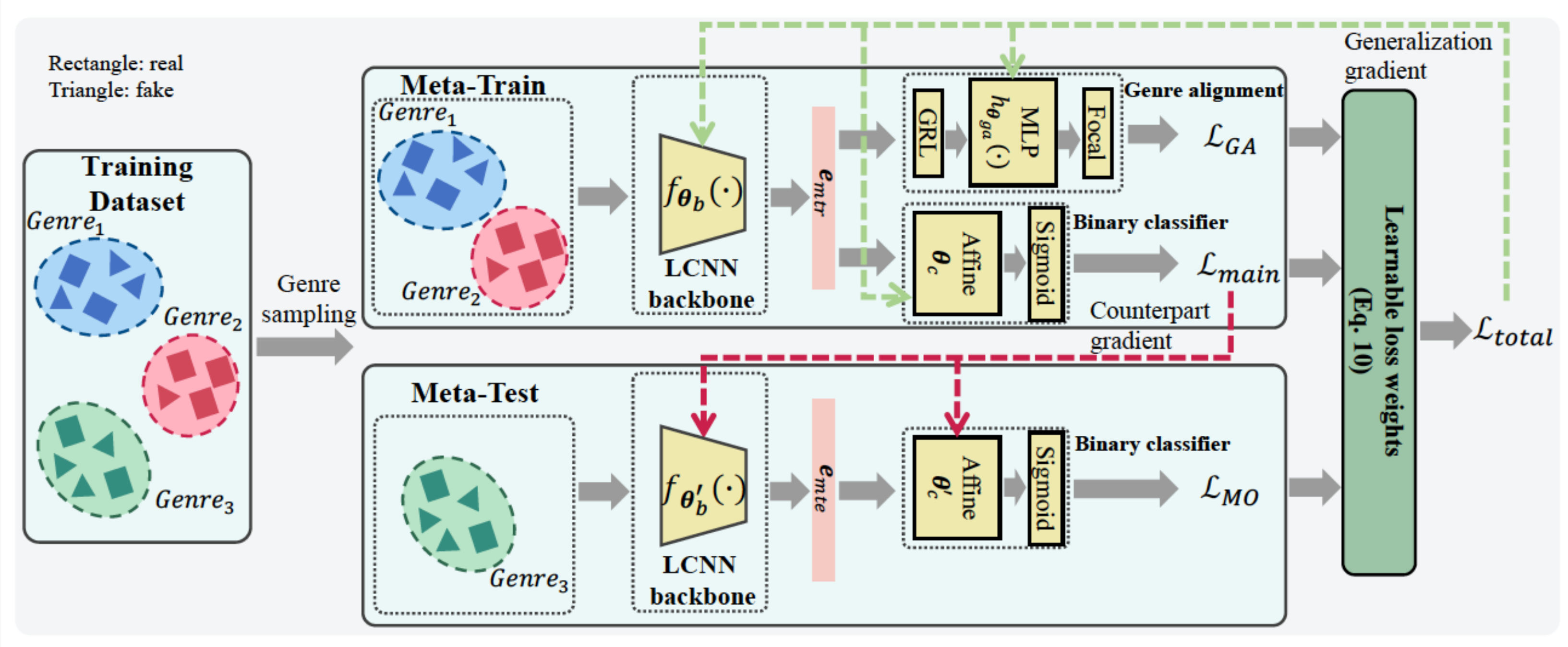
META-LEARNING PARADIGM FOR ANTI-SPOOFING TASK

- Bilevel optimization

$$\theta_{i+1} = \underset{\theta}{\operatorname{argmin}} \sum_{i}^{T} \mathcal{L}^{outer}(\theta_i^*, \mathcal{D}_{source}^{meta-test(i)}) + \mathcal{L}^{inner}(\theta_i, \mathcal{D}_{source}^{meta-train(i)})$$ On task sets

$$\text{s.t.} \quad \theta_i^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}^{inner}(\theta_i, \mathcal{D}_{source}^{meta-train(i)}),$$ On meta-train sets

- Anti-spoofing with meta-learning paradigm



$\mathcal{L}_{main} \rightarrow \mathcal{L}^{inner}$
$\mathcal{L}_{MO} \rightarrow \mathcal{L}^{outer}$
$\mathcal{L}_{GA} \rightarrow$ *Adversarial regularization term*

- Experimental results

Table 5.3: EER (%) of experimental results on CGP. For each protocol, the genre group in the bracket does not appear in the training dataset. A bold number means the best performance of this genre.

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)	$\mathcal{L}_{main}$	8.299									9.517	9.779
	$\mathcal{L}_{MO}, \mathcal{L}_{main}$	7.863										
	$\mathcal{L}_{GA}, \mathcal{L}_{main}$	8.238										
	$\mathcal{L}_{MO}, \mathcal{L}_{GA}, \mathcal{L}_{main}$	**7.511**										
CGP II (Group III)	$\mathcal{L}_{main}$	8.566									9.053	8.996
	$\mathcal{L}_{MO}, \mathcal{L}_{main}$	8.181										
	$\mathcal{L}_{GA}, \mathcal{L}_{main}$	8.481										
	$\mathcal{L}_{MO}, \mathcal{L}_{GA}, \mathcal{L}_{main}$	**7.764**										
CGP III (Group II)	$\mathcal{L}_{main}$	8.599									9.424	8.603
	$\mathcal{L}_{MO}, \mathcal{L}_{main}$	8.182										
	$\mathcal{L}_{GA}, \mathcal{L}_{main}$	8.505										
	$\mathcal{L}_{MO}, \mathcal{L}_{GA}, \mathcal{L}_{main}$	**8.032**										
CGP IV (Group I)	$\mathcal{L}_{main}$	8.160									8.966	8.657
	$\mathcal{L}_{MO}, \mathcal{L}_{main}$	7.827										
	$\mathcal{L}_{GA}, \mathcal{L}_{main}$	7.944										
	$\mathcal{L}_{MO}, \mathcal{L}_{GA}, \mathcal{L}_{main}$	**7.739**										

META-LEARNING PARADIGM FOR ANTI-SPOOFING TASK

- Summary
  - The meta-learning paradigm exhibits the potential to enhance the generalization capabilities of machine learning models.

CONTENTS

ISSUE 4: INTEGRATION

- We have addressed
  - Issue 1: Channel mismatch by pair-wise learning paradigm
  - Issue 2: Spoofing attacks by simulating spoofing attacks
  - Issue 3: Domain mismatch by meta-learning paradigm
- The remaining Issue 4
  - How can ASV systems be robust against these three threats simultaneously?

ISSUE 4: INTEGRATION

- Revisiting meta-learning paradigm from the bilevel optimization perspective

$$\theta_{i+1} = \underset{\theta}{\text{argmin}} \sum_{i}^{T} \mathcal{L}^{outer}(\theta_i^*, \mathcal{D}_{source}^{meta-test(i)}) + \mathcal{L}^{inner}(\theta_i, \mathcal{D}_{source}^{meta-train(i)})$$

$$\text{s.t.} \quad \theta_i^* = \underset{\theta}{\text{argmin}} \, \mathcal{L}^{inner}(\theta_i, \mathcal{D}_{source}^{meta-train(i)}),$$

→ Robust against domain mismatch

↓

Robust against channel mismatch and spoofing attacks?

↓

Spoofing-aware attention back-end!

↙ Simulation of spoofing attacks

↘ Pair-wise learning paradigm

DATA PREPARATION

- A new testing dataset is required. It should contain:
  - Channel mismatch scenario
  - Spoofing attacks scenario
  - Domain mismatch scenario
- CNComplex testing dataset based on the official CNCeleb testing dataset
  - The enrollment utterances remain unchanged
  - The evaluation utterances are subject to random substitution with re-vocoded data sourced from the CNSpoof dataset

spk1 utt1 target
spk1 utt2 nontarget
spk2 utt3 target
…
spkN uttM nontarget

Evaluation protocol
w/o spoofing attack

→

spk1 utt1-real target
spk1 utt2-real nontarget
spk2 utt3-HiFiGAN nontarget
…
spkN uttM-WORLD nontarget

Evaluation protocol
w/ spoofing attack

PRELIMINARY EXPERIMENT AND ANALYSIS

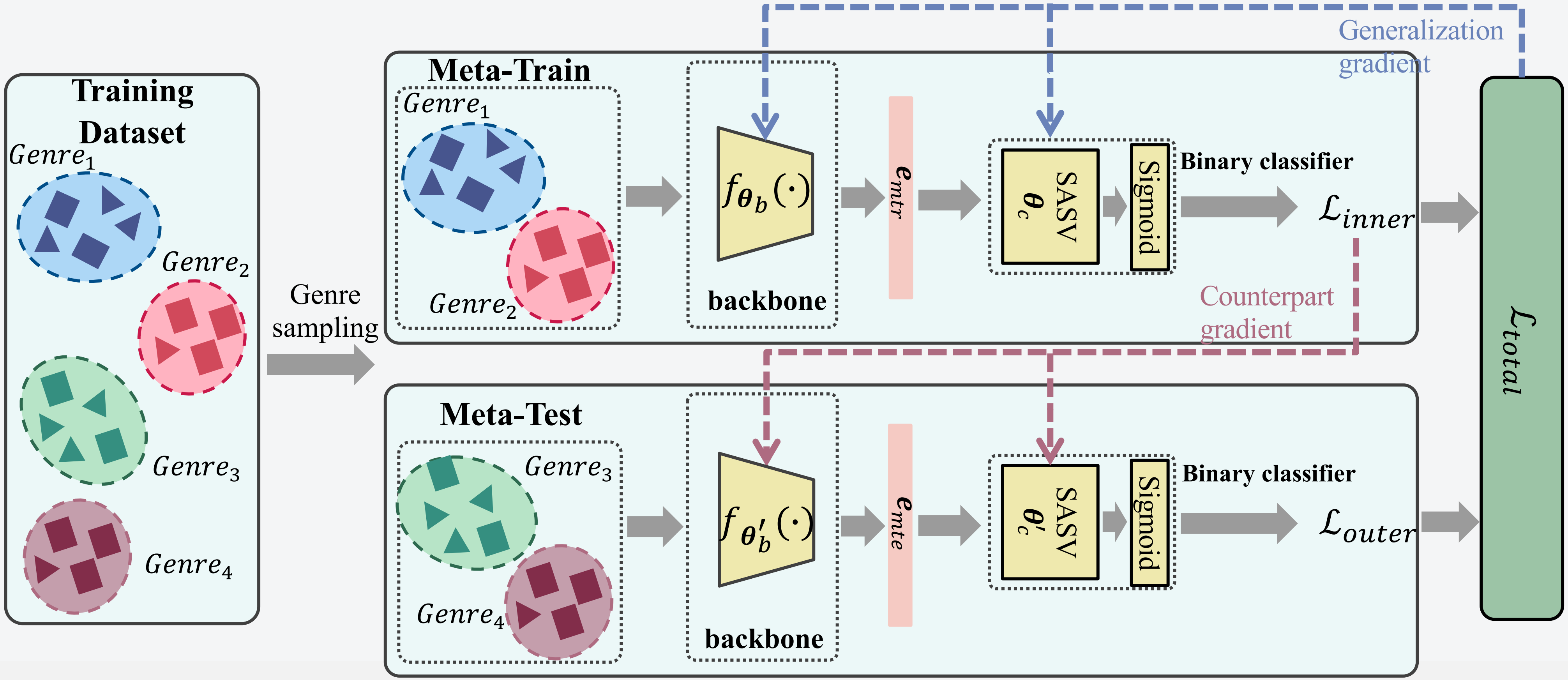
- Re-vocoded data can slightly improve the ASV performance
- SOTA ASV model is vulnerable when facing these threats

Table 6.1: We assess the performance of the ECAPA-TDNN model, trained on two distinct sets of data: CNCeleb 1&2 and CNCeleb 1&2 + CNSpoof according to cross-genre protocols. Subsequently, we conduct evaluations on both the original CNCeleb testing dataset and our newly developed testing dataset. The evaluation metrics employed for this assessment are the SV-EER and SASV-EER defined in Chapter 4.

Protocol	Training dataset	CNCeleb.Eval		CNComplex	
		SV-EER	SASV-EER	SV-EER	SASV-EER
CGP I	CNCeleb 1&2	9.38	-		37.64
	CNCeleb 1&2 + CNSpoof	9.02	-		36.97
CGP II	CNCeleb 1&2	10.04	-		40.76
	CNCeleb 1&2 + CNSpoof	9.75	-		40.17
CGP III	CNCeleb 1&2	10.12	-		39.62
	CNCeleb 1&2 + CNSpoof	9.33	-		38.49
CGP IV	CNCeleb 1&2	9.59	-		37.97
	CNCeleb 1&2 + CNSpoof	9.21	-		37.42

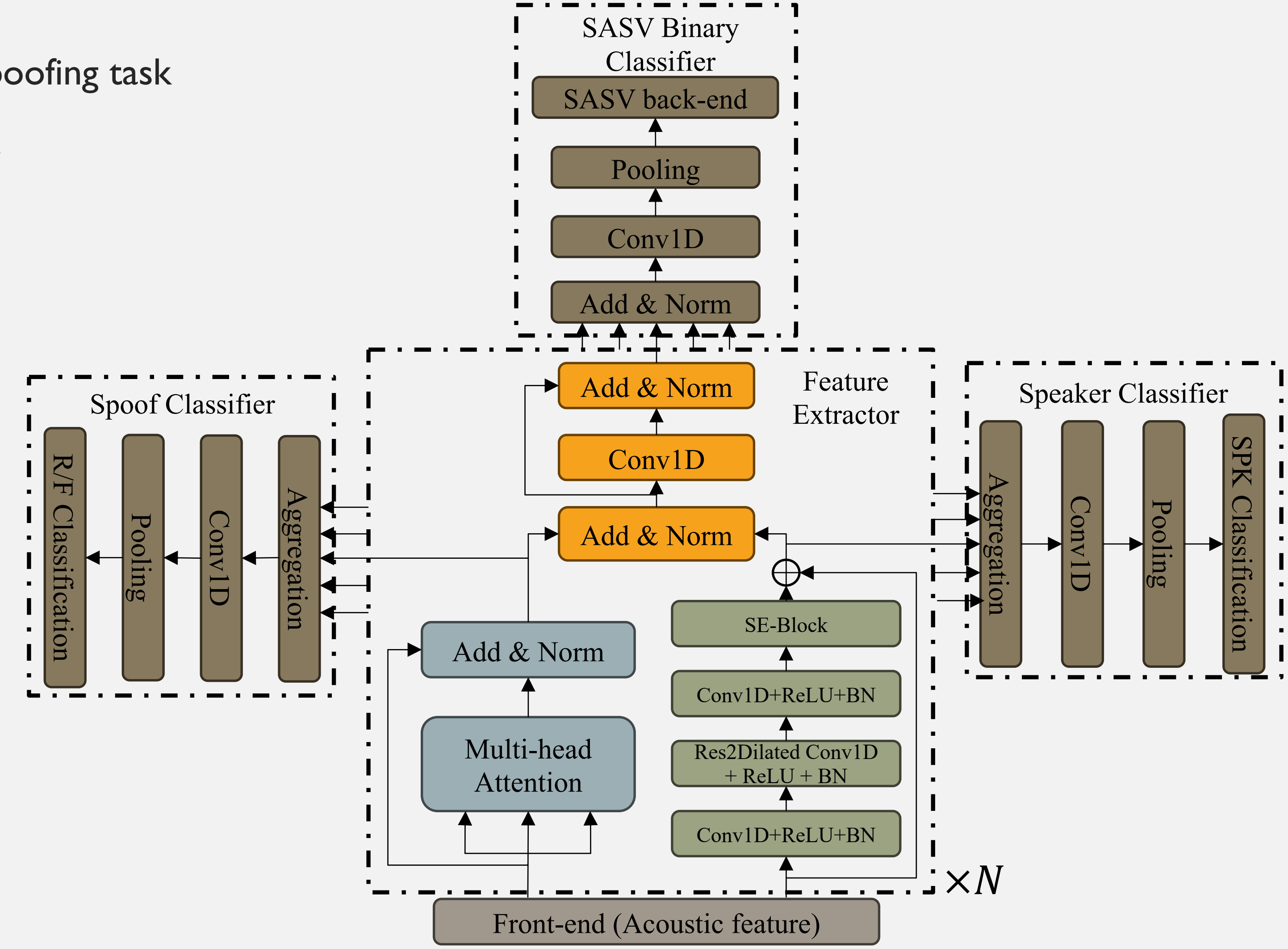
META-LEARNING PARADIGM: OUTER LOOP

- Backbone: asymmetric dual-path transformer-based model (illustrated on next page)



META-LEARNING PARADIGM: INNER LOOP

- Feature extractor: two branches
  - Attention block to extract information for anti-spoofing task
  - ECAPA block to extract information for ASV task
- Speaker classifier
- Spoof classifier
- SASV binary classifier
  - Pair-wise learning paradigm
  - Simulation of spoofing attacks



EXPERIMENTAL RESULTS AND ANALYSIS

- Robust against channel mismatch

dr: drama
en: entertainment
in: interview
lb: live broadcast
re: recitation
si: singing
sp: speech
vl: vlog

Table 6.2: The difference of EER. The first row of the table denotes genres of testing utterances, and the first column of the table represents genres of enrollment utterances. The blue number means our proposed approach is better than the baseline. The red number means our proposed approach is worse than the baseline.

	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

Evaluation genres

Enrollment genres

EXPERIMENTAL RESULTS AND ANALYSIS

- Robust against spoofing attacks

Table 6.3: The result of the proposed system on CNCeleb.Eval and CNComplex testing datasets. We evaluate the performance under the metrics of SV-EER and SASV-EER, respectively.

Protocol	Training dataset	CNCeleb.Eval		CNComplex	
		SV-EER	SASV-EER	SV-EER	SASV-EER
CGP I	CNCeleb 1&2	7.96	-		7.37
	CNCeleb 1&2 + CNSpoof	7.79	-		7.25
CGP II	CNCeleb 1&2	8.24	-		8.57
	CNCeleb 1&2 + CNSpoof	7.96	-		8.47
CGP III	CNCeleb 1&2	8.43	-		8.52
	CNCeleb 1&2 + CNSpoof	8.19	-		8.25
CGP IV	CNCeleb 1&2	8.23	-		7.73
	CNCeleb 1&2 + CNSpoof	8.13	-		7.48

EXPERIMENTAL RESULTS AND ANALYSIS

- Robust against domain mismatch

Table 6.4: 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 means the best performance of this genre.

Protocol	System	Overall	Group I			Group II			Group III		Group IV	
			dr	vl	sp	en	in	pl	lb	mo	si	re
CGP I	Baseline	21.31									28.63	14.58
(Group IV)	Our approach	17.42										
CGP II	Baseline	22.10									25.25	**13.17**
(Group III)	Our approach	18.48										
CGP III	Baseline	23.46									27.74	**13.10**
(Group II)	Our approach	20.02										
CGP IV	Baseline	22.59									24.87	11.33
(Group I)	Our approach	19.98										

CONTENTS

CONCLUSIONS AND FUTURE WORK

- Primary objectives
  - Improve the robustness of ASV systems to channel mismatch -> pair-wise learning paradigm
  - Improve the robustness of ASV systems to spoofing attacks -> simulation of spoofing attacks
  - Improve the robustness of ASV systems to domain mismatch -> meta-learning paradigm
  - Address these three threats jointly in an integrated manner -> incorporating pair-wise learning paradigm and spoofing attacks simulation into meta-learning paradigm

CONCLUSIONS AND FUTURE WORK

- Future work
  - Advanced architectures of neural network
    - Variants of the transformer model
  - Objective functions
    - Verification-based loss function vs Classification-based loss function
  - Learning paradigm
    - Test-time adaption, which is more flexible in the real-world setting
  - Self-supervised pretrained model
    - Transfer learning
  - Distillation of generative model
    - SpeechGPT
    - Audio foundation model

CONTENTS

PUBLICATIONS

- Speaker Verification (Thesis-related works)
  - **Chang Zeng**, Xin Wang, Erica Cooper, Xiaoxiao Miao, and Junichi Yamagishi, "Attention Back-End for Automatic Speaker Verification with Multiple Enrollment Utterances." ICASSP 2022.
  - **Chang Zeng**, Lin Zhang, Meng Liu, and Junichi Yamagishi, "Spoofing-Aware Attention based ASV Back-end with Multiple Enrollment Utterances and a Sampling Strategy for the SASV Challenge 2022." Proc. Interspeech 2022.
  - **Chang Zeng**, Xin Wang, Xiaoxiao Miao, Erica Cooper, and Junichi Yamagishi, "Improving Generalization Ability of Countermeasures for New Mismatch Scenario by Combining Multiple Advanced Regularization Terms." Proc. Interspeech 2023.
  - **Chang Zeng**, Xiaoxiao Miao, Xin Wang, Erica Cooper, and Junichi Yamagishi, "Joint Speaker Encoder and Neural Back-end Model for Fully End-to-End Automatic Speaker Verification with Multiple Enrollment Utterances." Computer Speech & Language.

PUBLICATIONS

- Singing Voice Synthesis
  - Chunhui Wang, **Chang Zeng (Co-first author)**, Jun Chen, and Xing He. "HiFi-WaveGAN: Generative Adversarial Network with Auxiliary Spectrogram-Phase Loss for High-Fidelity Singing Voice Generation." Submitted to ISNN 2024.
  - Chunhui Wang, **Chang Zeng (Co-first author)**, and Xing He. "XiaoiceSing 2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network." Proc. Interspeech 2023
  - Xintong Wang, **Chang Zeng (Co-first author)**, Jun Chen, and Chunhui Wang. "CrossSinger: A Cross Lingual Multi-Singer High-Fidelity Singing Voice Synthesizer Trained on Monolingual Singers." IEEE ASRU 2023
  - **Chang Zeng**, Chunhui Wang, and Xiaoxiao Miao. "InstructSing: High-fidelity Singing Voice Generation via Instructing Yourself." Submitted to ICME 2024

PUBLICATIONS

- Others
  - Weixin Zhu, Zilin Wang, Jiuxin Lin, **Chang Zeng**, and Tao Yu, "SSI-Net: A Multi-Stage Speech Signal Improvement System for ICASSP 2023 SSI Challenge," ICASSP 2023.
  - Meng Liu, Kong Aik Lee, Longbiao Wang, Hanyi Zhang, **Chang Zeng**, and Jianwu Dang, "Cross-Modal Audio-Visual Co-Learning for Text-Independent Speaker Verification," ICASSP 2023.
  - Kai Li, Sheng, Li, Xugang Lu, Masato Akagi, Meng Liu, Lin Zhang, **Chang Zeng**, Longbiao Wang, Jianwu Dang, and Masashi Unoki. "Data Augmentation Using McAdams Coefficient-Based Speaker Anonymization for Fake Audio Detection," Proc. Interspeech 2022.
  - Xiaohui Liu, Meng Liu, Lin Zhang, Linjuan Zhang, **Chang Zeng**, Kai Li, Nan Li, Kong Aik Lee, Longbiao Wang, and Jianwu Dang. "Deep Spectro-temporal Artifacts for Detecting Synthesized Speech," Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. 2022.
  - Meng Liu, Longbiao Wang, Kong Aik Lee, Hanyi Zhang, **Chang Zeng**, and Jianwu Dang, "DeepLip: A Benchmark for Deep Learning-Based Audio-Visual Lip Biometrics," 2021 ASRU.

- XiaoiceSing2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network

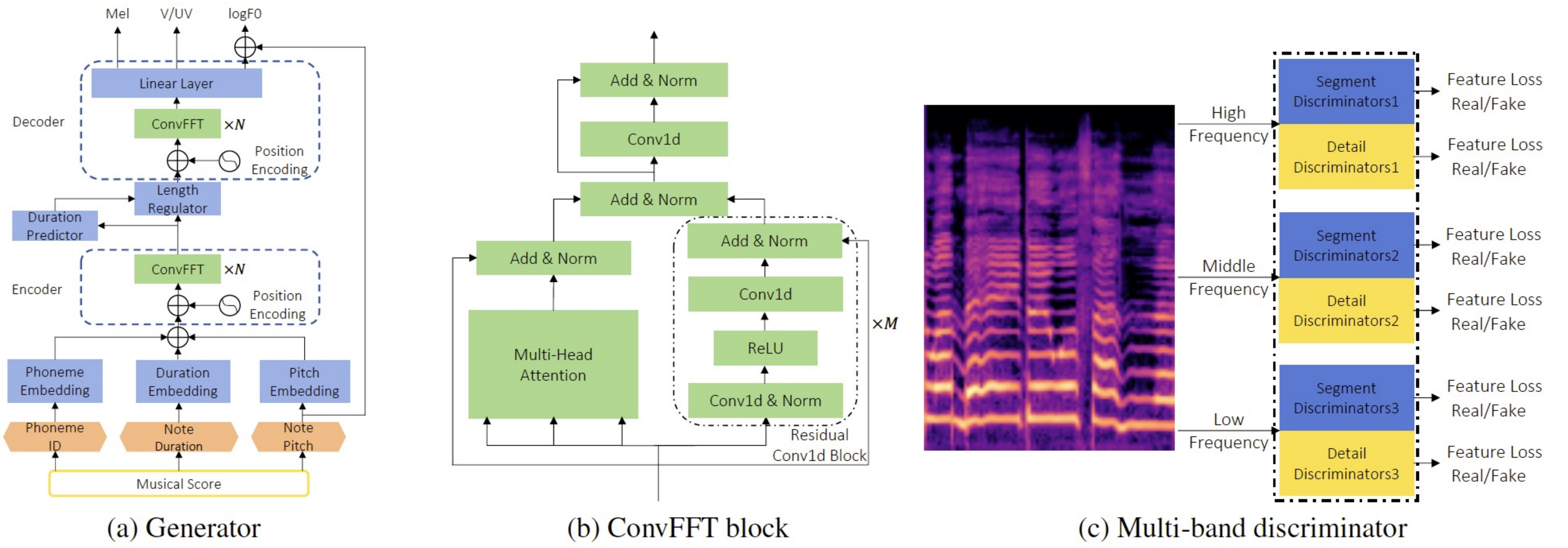


Figure 1: *The architecture of XiaoiceSing2. (a). The improved feed-forward Transformer. (b). Feed-forward Transformer with parallel residual convolutional block. (c). Multi-band discriminator, consisting of three sub-discriminators, and each contains several segment discriminators and detail discriminators.*

- CrossSinger: A Cross-Lingual Multi-Singer High-Fidelity Singing Voice Synthesizer Trained on Monolingual Singers

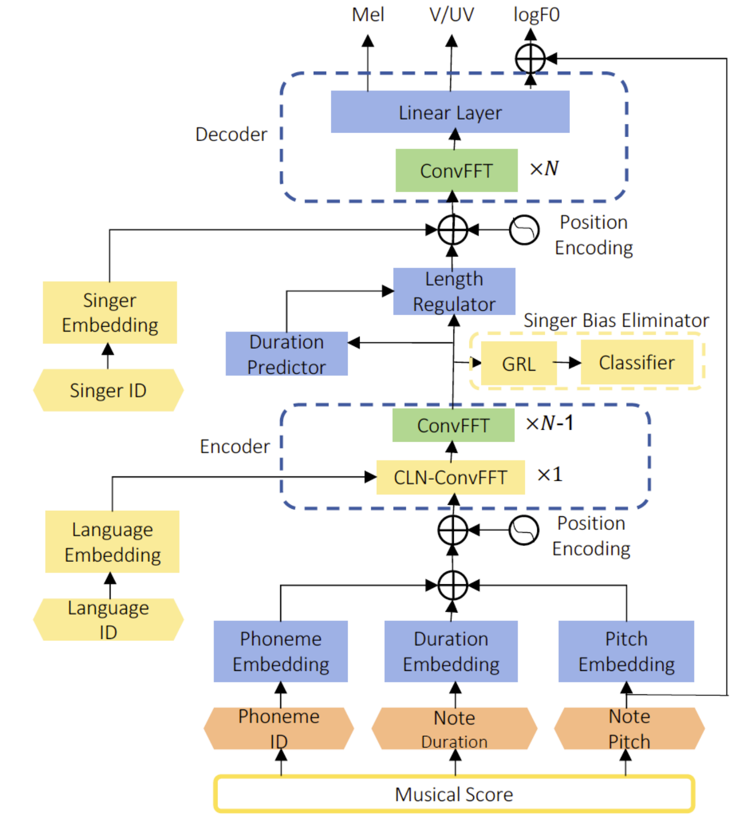


Fig. 1. The architecture of the CrossSinger generator. Yellow blocks denote the improvement parts compared with Xiaoicesing2.

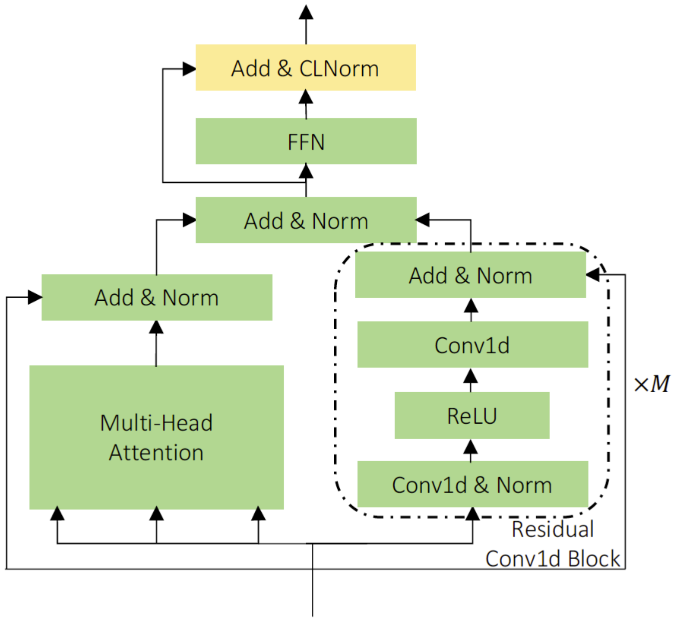


Fig. 2. The architecture of ConvFFT block with conditional layer normalization. CLN is used to replace the last layer normalization of ConvFFT to introduce the language information.

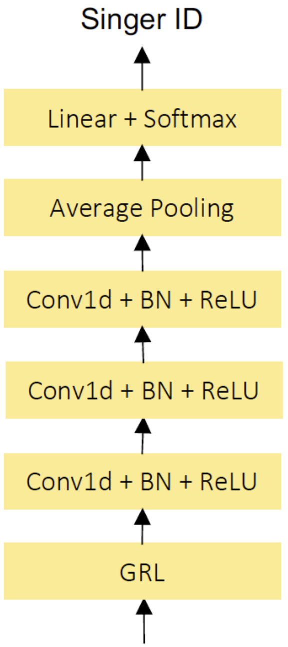


Fig. 3. The architecture of singer bias eliminator. It is utilized to remove singer biases implicitly associated with lyrics.

Thanks for attention