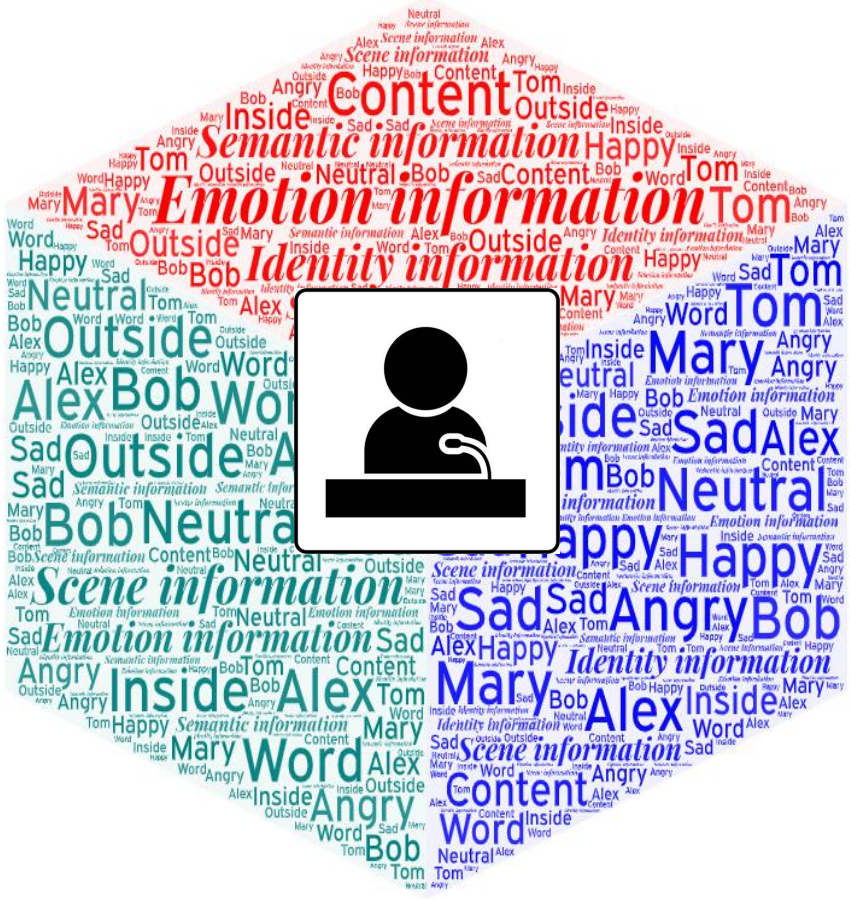
SEEF-ALDR: A Speaker Embedding Enhancement Framework via Adversarial Learning based Disentangled Representation

Jianwei Tai, Xiaoqi Jia, Qingjia Huang, Weijuan Zhang, Haichao Du, Shengzhi Zhang

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

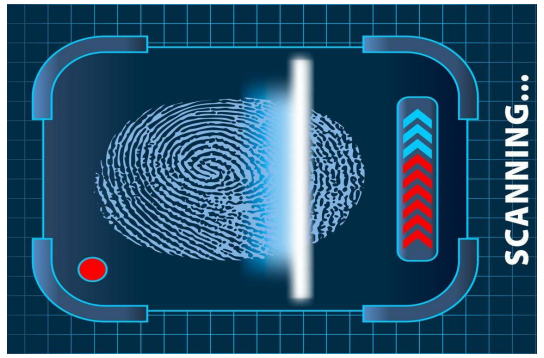
中国科学院大学
University of Chinese Academy of Sciences

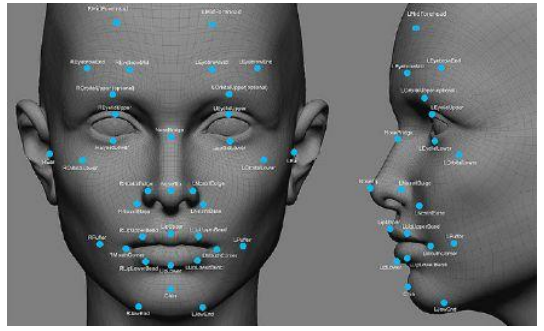
Idea

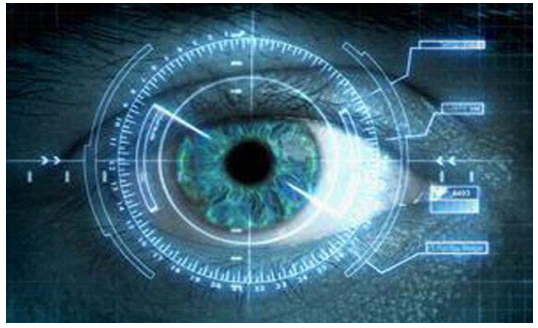

Emotion information
Semantic information
Identity information
Scene information
Content
Speakers
Decoupling
Emotion
Text
Scenes
Identity Features
Speaker Verification Task
Identity-unrelated Features

Motivation

1. Biometric authentication relies on biosensors to collect physiological and behavioral characteristics of users and applies the biostatistics principle to verify users' identity.

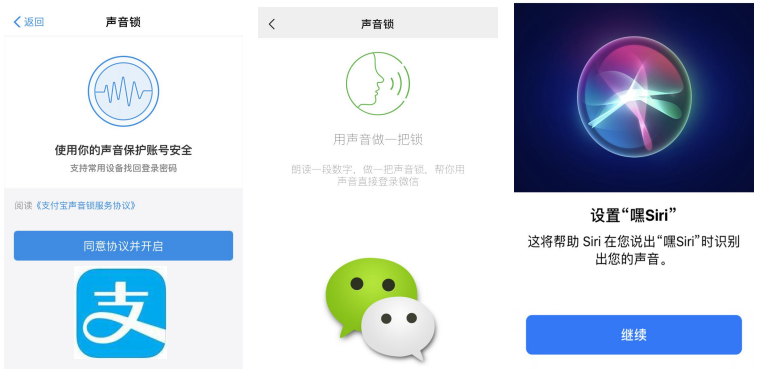






Motivation

2 As the advance of machine learning and the pervasiveness of mobile devices, speaker verification has become a topic of interest for its applications in authentication.



Motivation

3 The task of ``in-the-wild'' speaker verification is still quite challenging.

Who is the speaker?

What was said?

What is the scene of dialogue?

In what mood?

Related Works

1. Existing works directly process the original speech, which contains lots of unrelated features other than the ones closely to the speaker identity.

2. Existing works usually require additional supervision during the training process and encode each attribute as a separate element in feature vectors, which may not be easy in real-world situations.

3. Existing works do not generalize well, so it is difficult to port existing models to improve the performance of speaker verification.

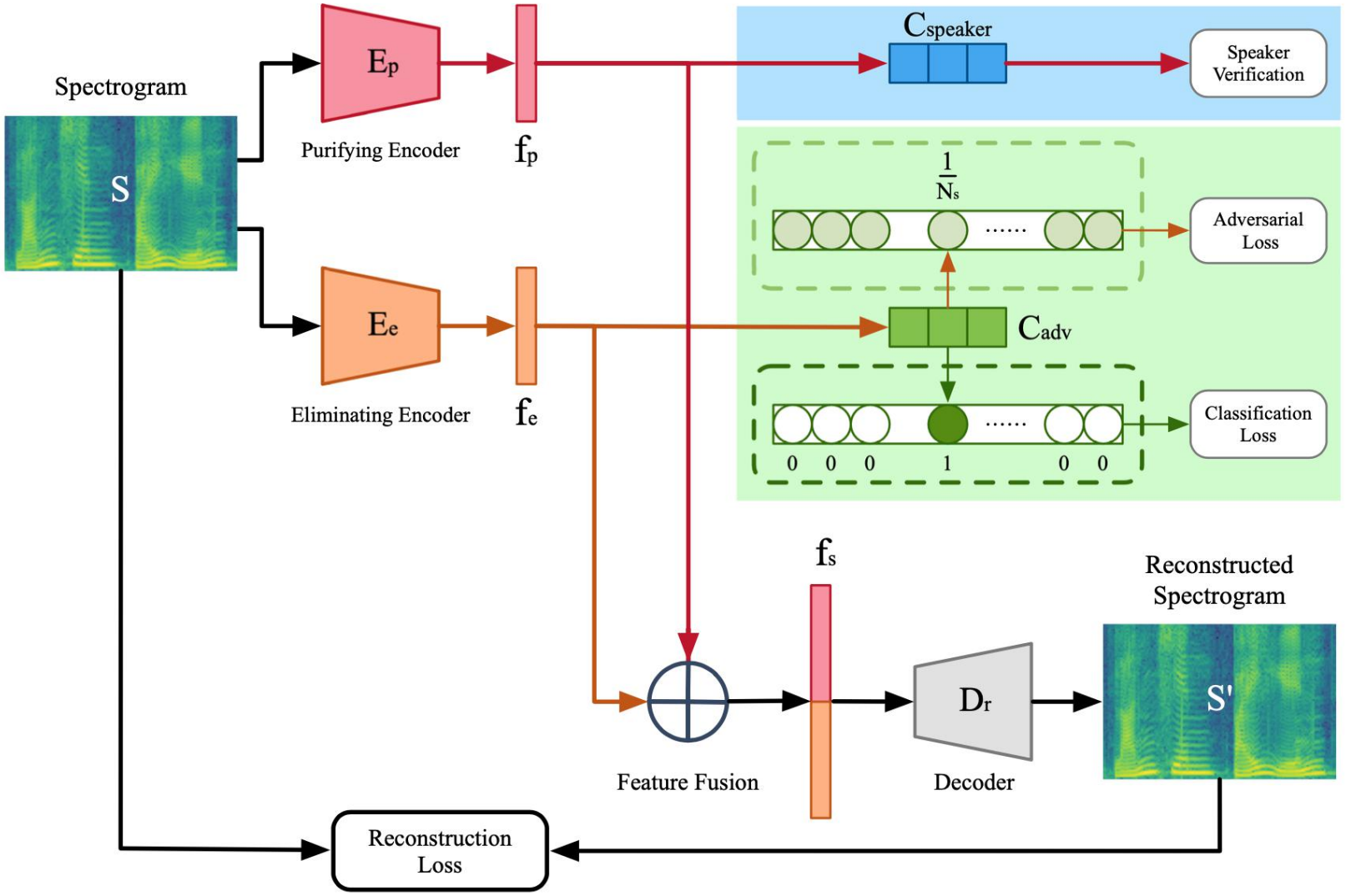
Our Contributions

1. We propose a novel twin network-based speaker embedding enhancement framework to significantly improve the performance of speaker verification by combining the autoencoder-like architecture and adversarial learning.

2. Our framework follows the modular design to easy the porting of existing models without adjusting the structure of them.

3. We only use speaker labels to train the eliminating encoder based on adversarial supervision to obtain the identity-unrelated information without additional labels.

SEEF-ALDR



- **E_p** → Extracting identity-related information
- **E_e** → Extracting identity-unrelated information
- **F_p** → Speaker embedding
- **F_e** → Feature embedding complementary to speaker
- **F_s** → Embedding of Fused features
- **D_r** → For reconstruction of original speech
- **$C_{speaker}$** → For speaker classification task
- **C_{adv}** → Decoupling speaker and identity-unrelated features by adversarial learning

SEEF-ALDR

- Our framework tries to decouple the identity-related information and identity-unrelated information from the input speech signal by using adversarial learning.

- After training process, Our framework can learn the distribution of the identity-related information and identity-unrelated information in the feature space from the speech samples and decouple them confidently.

- To improve the portability, we follow a modular design. The encoders, classifiers, feature fusion, and the reconstruction decoder are all implemented in the form of modules.

Experiments

1 Datasets

- In our experiments, we only use audio files from them for speaker verification task.
- Both of them are fairly gender-balanced, and speakers in them span a wide range of races, professions, ages, emotions, accents and so on.

Table 1: Training and Testing Dataset on Voxceleb1.

Dataset	Training	Testing	Total
#POIs	1211	40	1251
#Utterances	148,642	4,874	153,516

Table 2: Training and Testing Dataset on Voxceleb2.

Dataset	Training	Testing	Total
#POIs	5994	118	6112
#Utterances	1,092,009	36,237	1,128,246

Experiments

2 Performance of SEEF-ALDR

- We chose state-of-the-art speaker verification models as baselines to port into our framework and evaluate the effectiveness of it.
- When reproducing those models, we ensure that the model structure, loss function, test dataset, and similarity metric are consistent with those in the original paper.
- After porting them into our framework, we retrain it on Vox1 and Vox2 respectively.
- We choose two metrics: the detection cost function and the Equal Error Rate (EER), to evaluate the system performance.

Experiments

2 Performance of SEEF-ALDR on Vox1

	Model	Loss Function	Dims	Aggregation	Metric	EER (%)	C_{det}	EER-IP
Nagrani et al. [46]	VGG-M	Softmax	512	TAP	Cosine	7.8	0.71	-
SEEF-ALDR	VGG-M	Softmax	512	TAP	Cosine	**6.51**	**0.619**	**16.7%**
Li et al. [32]	ResCNN	Softmax+Triplet	512	TAP	Cosine	4.80	N/A	-
SEEF-ALDR	ResCNN	Softmax+Triplet	512	TAP	Cosine	**4.02**	**0.469**	**16.3%**
Bhattacharya et al. [1]	VGGnet	Softmax	512	TAP	PLDA	4.52	N/A	-
SEEF-ALDR	VGGnet	Softmax	512	TAP	PLDA	**3.95**	**0.439**	**12.6%**
Cai et al. [4]	ResNet-34	Softmax	N/A	TAP	Cosine	5.48	0.553	-
SEEF-ALDR	ResNet-34	Softmax	256	TAP	Cosine	**4.31**	**0.454**	**21.4%**
Cai et al. [4]	ResNet-34	Softmax	N/A	TAP	PLDA	5.21	0.545	-
SEEF-ALDR	ResNet-34	Softmax	256	TAP	PLDA	**4.35**	**0.479**	**16.5%**
Cai et al. [4]	ResNet-34	A-Softmax	N/A	TAP	Cosine	5.27	0.439	-
SEEF-ALDR	ResNet-34	A-Softmax	256	TAP	Cosine	**4.26**	**0.433**	**19.2%**
Cai et al. [4]	ResNet-50	Softmax	N/A	SAP	Cosine	5.51	0.522	-
SEEF-ALDR	ResNet-50	Softmax	256	SAP	Cosine	**4.40**	**0.469**	**20.1%**
Cai et al. [4]	ResNet-50	A-Softmax	N/A	SAP	Cosine	4.90	0.509	-
SEEF-ALDR	ResNet-50	A-Softmax	256	SAP	Cosine	**4.08**	**0.455**	**16.7%**
Hajibabaei et al. [14]	ResNet-20	Softmax	256	TAP	Cosine	6.98	0.540	-
SEEF-ALDR	ResNet-20	Softmax	256	TAP	Cosine	**4.56**	**0.503**	**34.7%**
Hajibabaei et al. [14]	ResNet-20	Softmax	128	TAP	Cosine	6.73	0.526	-
SEEF-ALDR	ResNet-20	Softmax	128	TAP	Cosine	**4.48**	**0.497**	**33.4%**
Hajibabaei et al. [14]	ResNet-20	Softmax	64	TAP	Cosine	6.31	0.527	-
SEEF-ALDR	ResNet-20	Softmax	64	TAP	Cosine	**4.37**	**0.494**	**30.7%**
Hajibabaei et al. [14]	ResNet-20	A-Softmax	128	TAP	Cosine	4.40	0.451	-
SEEF-ALDR	ResNet-20	A-Softmax	128	TAP	Cosine	**3.81**	**0.437**	**13.4%**
Hajibabaei et al. [14]	ResNet-20	A-Softmax	64	TAP	Cosine	4.29	0.442	-
SEEF-ALDR	ResNet-20	A-Softmax	64	TAP	Cosine	**3.62**	**0.437**	**15.6%**

Experiments

2 Performance of SEEF-ALDR on Vox2

	Model	Loss Function	Test set	EER (%)	C_{det}	EER-IP
Xie et al. [68]	Thin ResNet+NV	Softmax	Voxceleb1	3.57	N/A	-
SEEF-ALDR	Thin ResNet+NV	Softmax	Voxceleb1	**2.85**	**0.327**	**20.2%**
Xie et al. [68]	Thin ResNet+GV	Softmax	Voxceleb1	3.22	N/A	-
SEEF-ALDR	Thin ResNet+GV	Softmax	Voxceleb1	**2.61**	**0.335**	**19.0%**
Xie et al. [68]	Thin ResNet+NV	Softmax	Voxceleb1-E	3.24	N/A	-
SEEF-ALDR	Thin ResNet+NV	Softmax	Voxceleb1-E	**2.87**	**0.373**	**11.4%**
Xie et al. [68]	Thin ResNet+GV	Softmax	Voxceleb1-H	5.17	N/A	-
SEEF-ALDR	Thin ResNet+GV	Softmax	Voxceleb1-H	**4.52**	**0.520**	**12.6%**
Chung et al. [6]	ResNet-34	Softmax + Contrastive	Voxceleb1	5.04	0.543	-
SEEF-ALDR	ResNet-34	Softmax + Contrastive	Voxceleb1	**3.08**	**0.334**	**38.8%**
Chung et al. [6]	ResNet-50	Softmax + Contrastive	Voxceleb1	4.19	0.449	-
SEEF-ALDR	ResNet-50	Softmax + Contrastive	Voxceleb1	**2.75**	**0.326**	**34.4%**
Chung et al. [6]	ResNet-50	Softmax + Contrastive	Voxceleb1-E	4.42	0.524	-
SEEF-ALDR	ResNet-50	Softmax + Contrastive	Voxceleb1-E	**3.25**	**0.398**	**26.5%**
Chung et al. [6]	ResNet-50	Softmax + Contrastive	Voxceleb1-H	7.33	0.673	-
SEEF-ALDR	ResNet-50	Softmax + Contrastive	Voxceleb1-H	**5.30**	**0.575**	**27.7%**

Experiments

3 Performance of Disentangled Representation

- To demonstrate the difference between the two decoupled features from SEEF-ALDR, we to reduce the dimension of the high-level features and visualize them by T-SNE.

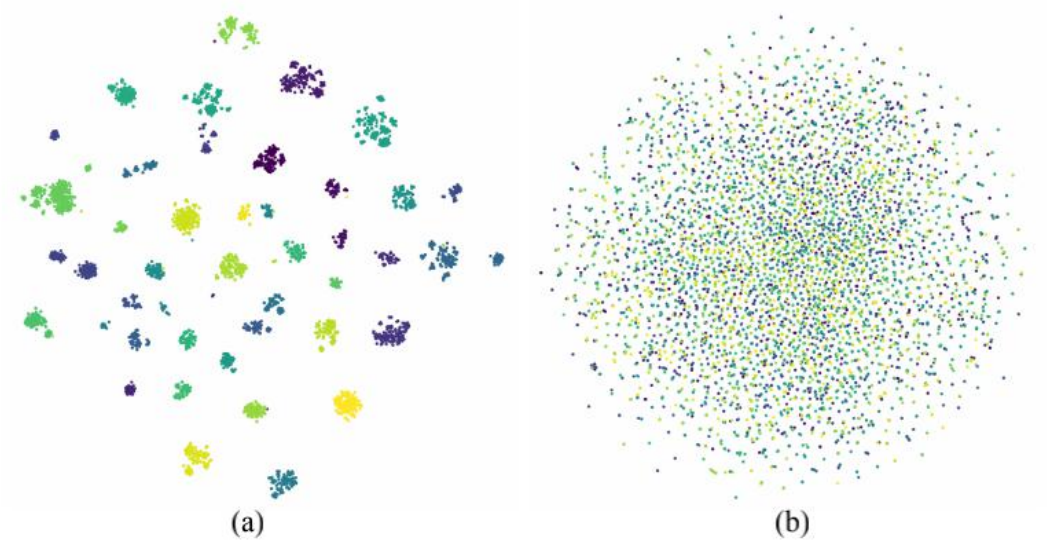


Figure 3: T-SNE Visualization of the Decoupled Features based on the test set of Voxceleb1: (a) Features Extracted by E_p (b) Features Extracted by E_e.

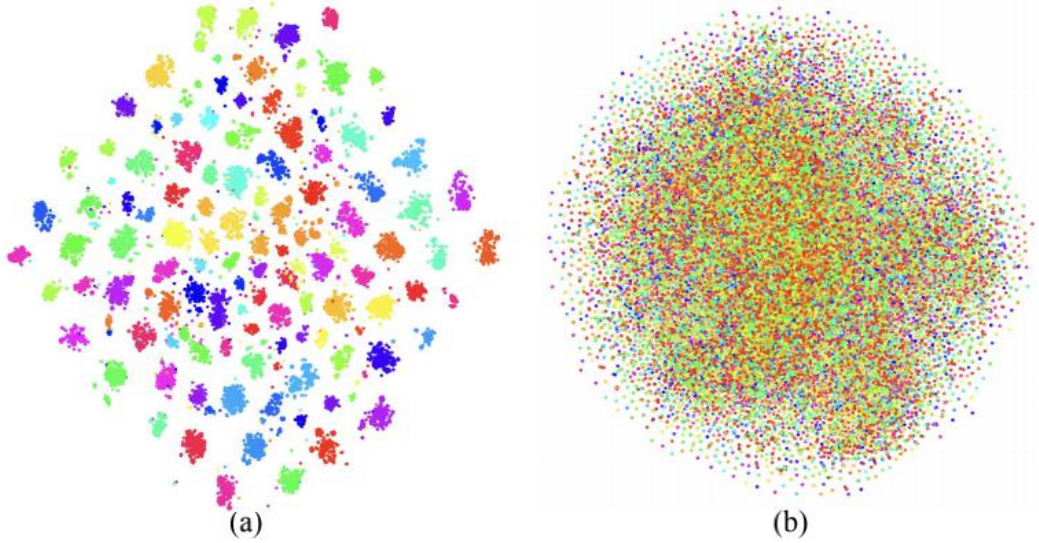


Figure 4: T-SNE Visualization of the Decoupled Features based on the test set of Voxceleb2: (a) Features Extracted by E_p (b) Features Extracted by E_e.

Experiments

4 Ablation Study

- To further evaluate the contribution of each module in SEEF-ALDR, the ablation study has been conducted.
- We use Vox2 as the training dataset, Vox1 as the testing dataset, and the same metrics, the detection cost function and the Equal Error Rate as in the above experiments to evaluate the performance of speaker verification.

Branch	EER (%)	C_{det}	EER-IP
E_p	5.04	0.543	-
$E_p + D_r$	4.42	0.509	12.3%
E_e	49.79	0.999	-887.9%
E_e w/o L_s^{adv}	30.46	0.999	-504.4%
E_e w/o L_e^{adv}	34.57	0.999	-585.9%
$E_p + RandomVector + D_r$	4.35	0.511	13.6%
SEEF-ALDR($E_p + E_e + D_r$)	**3.08**	**0.334**	**38.8%**

Discussion

1. Novel identity impersonation attack based on SEEF-ALDR:
New speeches can be created by the reconstruction decoder based on the cross-fusion the two decoupled features of them, e.g., identity features of Speaker A and identity-unrelated features of Speaker B, and vice versa. Then the identity impersonation attack can be executed.

2. Better sound event detection based on SEEF-ALDR:
It is a novel way to accurately extract the sound of interest from the information-rich audio to identify the event.

Conclusion

1. We propose a novel speaker embedding enhancement framework via adversarial learning based disentangled representation, to decouple the speaker identity features and the identity-unrelated ones from original speech.

2. Experiment results demonstrate that our framework can construct more accurate speaker embeddings for existing speaker verification models to improve the performance of “in-the-wild” speaker verification with little effort.

Q & A

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

Thank You!