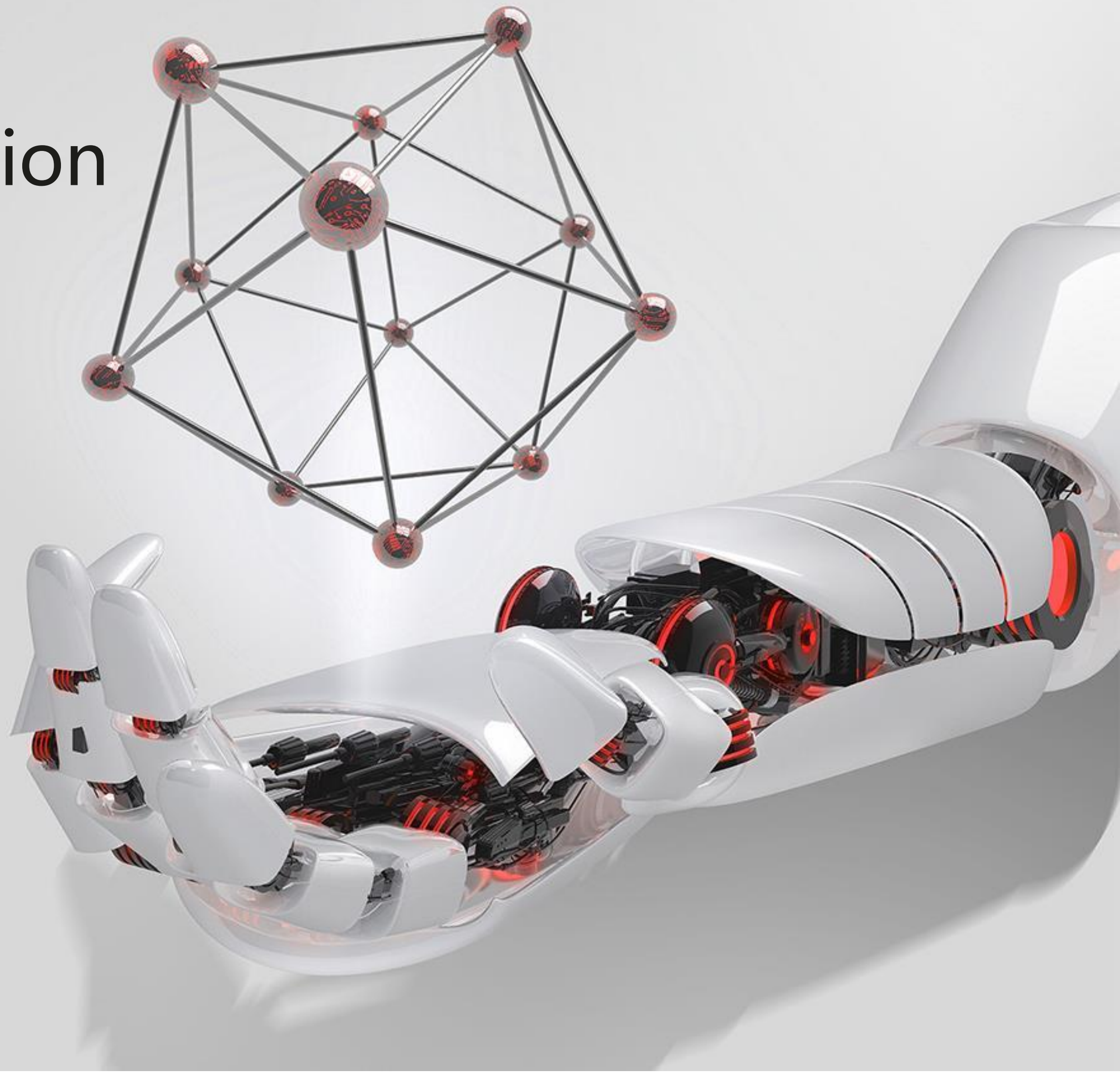
Text-Independent Speaker Verification with Adversarial Learning on Short Utterances



Authors: Kai Liu, Huan Zhou

Presented by: Kai Liu, Huan Zhou

IEEE Signal Processing Society
IEEE Advancing Technology for Humanity
ICASSP2020 Barcelona
HUAWEI

Content

Introduction

Speaker Verification System

- i-vector
- X-vector
- D-vector
- G-vector

  …

Short-utterance Speaker Verification

- Performance decline dramatically

  e.g. (NIST-SRE 2010) i-vector/PLDA EER :

  2.48%(full) → 24.78%(5 seconds)

Introduction

Improvement

- feature extraction techniques, intermediate parameter estimation, speaker model generation, score normalization
- teacher-student framework & scoring scheme calibration
- duration robust speaker embeddings
    - NN architectures: Inception Net, Inception-ResNet, ResCNN, GANs, …
    - Losses: triplet loss, am-softmax, …

Related Works

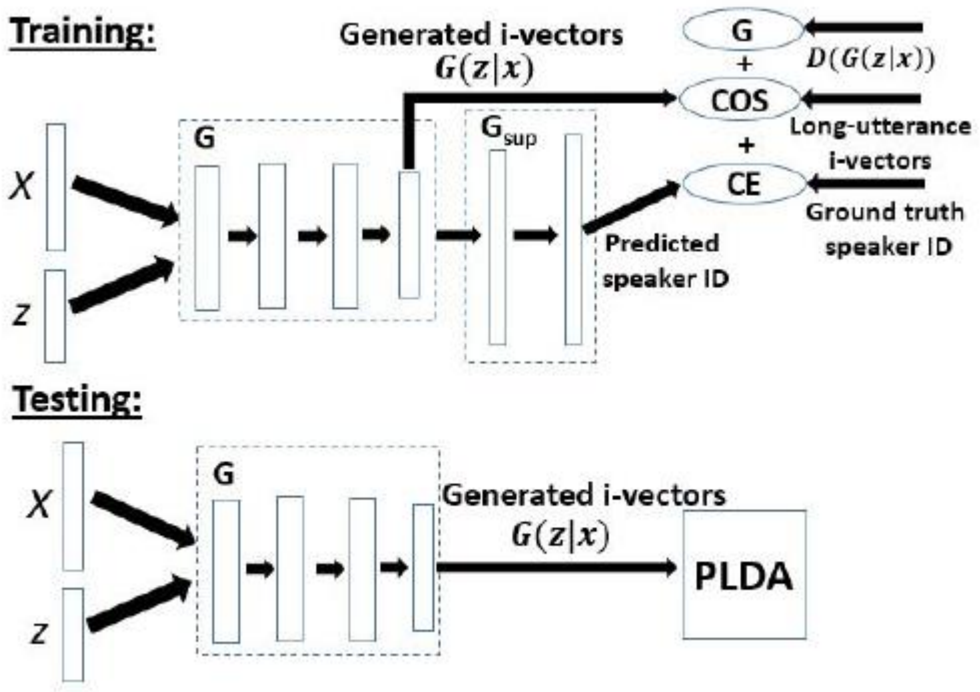


Figure 3: *Training of the generator network G and its application in the testing stage.*

Table 1: *The speaker verification results in terms of EER (%) on all the three conditions of the SRE08 "short2-10sec" male trail list.*

	EER (%)			
System	**Cond. 6**	**Cond. 7**	**Cond. 8**	**Average**
a) Baseline	7.28	6.15	6.06	6.50
b) Single G	10.04	8.85	8.33	9.07
c) a + b	7.28	5.77	6.06	6.37
d) D-WCGAN	9.45	8.08	8.33	8.62
e) a + d	**6.89**	**5.77**	**5.30**	**5.99**

Table 2: *The speaker verification results in terms of EER (%) on all the three conditions of the SRE08 "10sec-10sec" male trail list.*

	EER (%)			
System	**Cond. 6**	**Cond. 7**	**Cond. 8**	**Average**
a) Baseline	11.97	10.32	9.60	10.63
b) Single G	15.32	13.89	12.00	13.77
c) a + b	11.16	10.71	9.60	10.49
d) D-WCGAN	15.42	13.89	13.60	14.30
e) a + d	**10.75**	**8.73**	**8.80**	**9.43**

cite: Ivector transformation using conditional generative adversarial networks for short utterance speaker verification

Proposed Approach

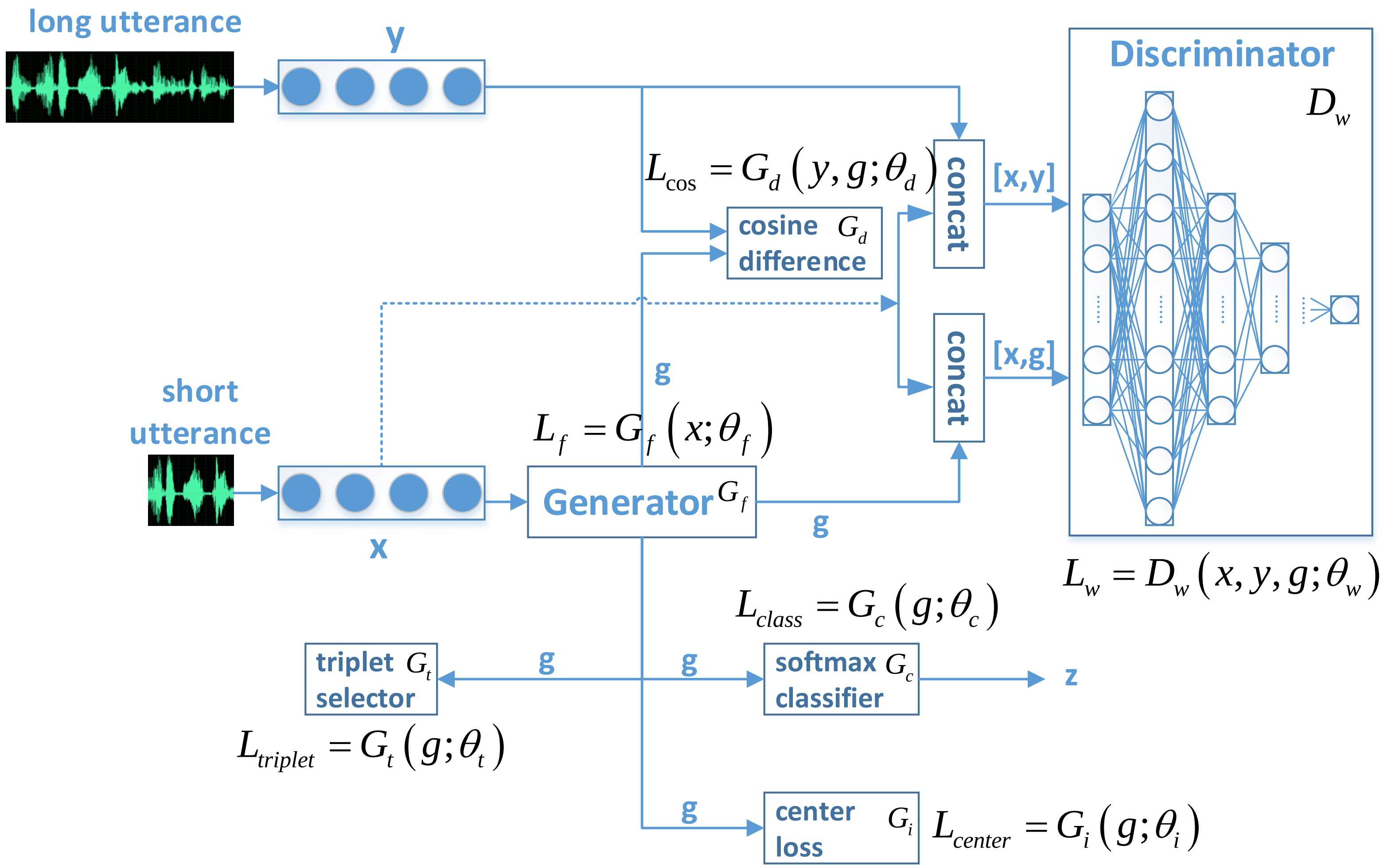


Fig. 1.1. Framework of our proposed system

Proposed Approach

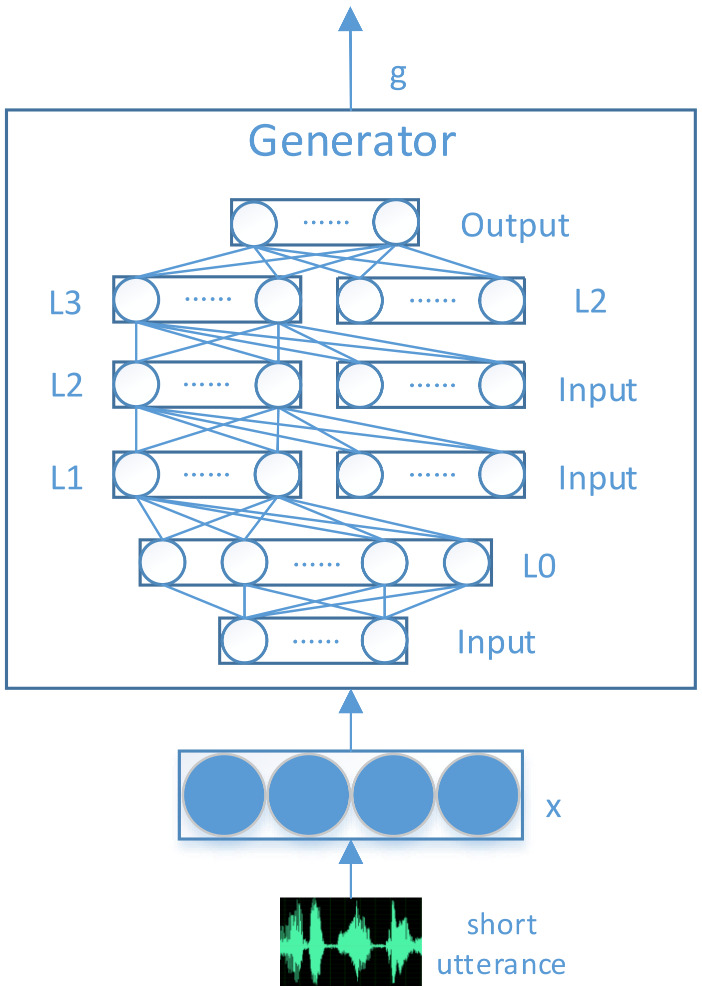


Fig. 1.2. Generator network structure

Proposed Approach

Discriminator-Related Loss Functions

- conditional wasserstein distance loss

$$\min_{G_f} \max_{D_w} L_{cw}(D_w, G_f) = E_y[D_w(y;x)] + E_x[D_w(G_f(x);x))]$$

- Fr´echet Inception Distance (fid) loss

$$L_{fid} = |\mu_y - \mu_g|^2 + tr\left(C_y + C_g - 2(C_y C_g)^{\frac{1}{2}}\right)$$

Proposed Approach

Generator-Related Loss Functions

- softmax loss

$$L_{class} = \frac{1}{N}\sum_{i=1}^{N} log \frac{e^{W_{z_i}^T g_i + b_{z_i}}}{\sum_{j=1}^{c} e^{W_j^T g_i + b_j}}$$

- triplet loss

$$L_{triplet} = \sum_{\gamma \in \Gamma} max\left(\| g_a - g_p \|_2^2 - \| g_a - g_n \|_2^2 + \Psi, 0\right)$$

- center loss

$$L_{center} = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2$$

- cosine loss

$$L_{cos} = 1 - \bar{g}^* \bar{y}$$

Proposed Approach

Total Loss Functions

- Discriminator

$$L_W = L_w / L_{cw} + \lambda L_{fid}$$

- Generator

$$L_G = L_w / L_{cw} + \alpha L_{class} + \beta L_{cos} + L_{center} + \epsilon L_{triplet}$$

Dataset

Train Set

- subset of voxceleb2
- 1,057 speakers
- 164,716 utterances (randomly cut to 2 seconds vs. original wav)

Test Set

- subset of voxceleb1
- 40 speakers
- 13,265 utterance pairs (randomly cut to 2 seconds and 1 second)

Experiments

Table 1. System descriptions

system	L_c	L_{cos}	L_t	L_{class}	L_{cw}	L_{fid}
v1	√	√		√	√	√
v2	√	√		√	√	
v3			√ a	√	√	
v4			√ a	√		
v6		√	√ b	√	√	
v5			√ a		√	
v7	√	√	√ b	√	√	
v8			√ b	√	√	

L_t : a means that inputs are sampled from both y and g and b means from g only

Experiments

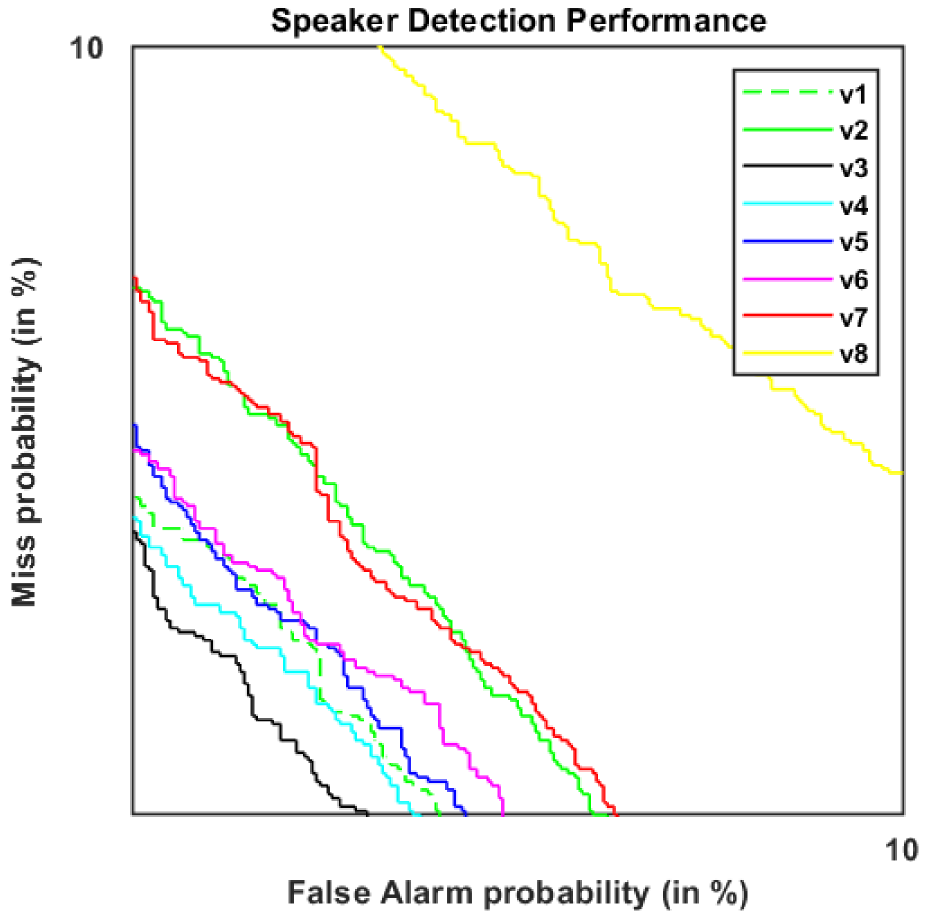
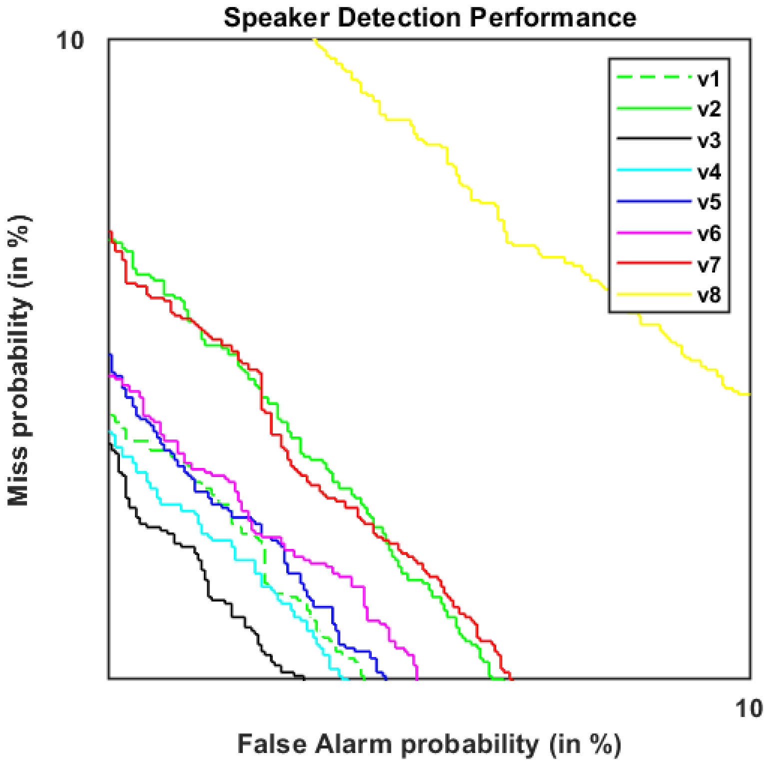


Fig. 2. DET performances for different systems

ps : we compute EER by compare embedding cosine distance

Experiments

- FID loss has positive effect (v1 vs. v2);
- Conditional WGAN outperforms WGAN (v3 vs. v4);
- Triplet loss is preferred (v7 vs. v2);
- Triplet a greatly outperforms triplet b (v3 vs. v8);
- softmax has positive effect (v3 vs. v5);
- Center loss has negative effect (v6 vs. v7);
- Cosine loss has significant positive effect (v6 vs. v8).



system	L_c	L_{cos}	L_t	L_{class}	L_{cw}	L_{fid}
v1	√	√		√	√	√
v2	√	√		√	√	
v3			√ a	√	√	
v4			√ a	√		
v6		√	√ b	√	√	
v5			√ a		√	
v7	√	√	√ b	√	√	
v8			√ b	√	√	

Experiments

Table 2. Comparison with the baseline system

system	2s-2s		1s-1s	
	EER(%)	minDCF	EER(%)	minDCF
G-vector	7.557	0.8170	14.133	0.866
ours	7.237	0.7578	13.599	0.881
fusion	7.168	0.7734	13.400	0.866

Conclusion

- proposed enhanced embedding for short-utterance speaker verification with Wasserstein Conditional GAN
- validated the effectiveness of a bunch of loss criteria on the GAN training

Future work

- better GAN structure
- more data
- how to describe distribution similarity in a better way
- GAN inside embedding extraction network
- more training tricks

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

HUAWEI