TEXT-INDEPENDENT SPEAKER VERIFICATION WITH ADVERSARIAL LEARNING ON SHORT UTTERANCES

Kai Liu, Huan Zhou

Artificial Intelligence Application Research Center, Huawei Technologies
Shenzhen, PRC

ABSTRACT

A text-independent speaker verification system suffers severe performance degradation under short utterance condition. To address the problem, in this paper, we propose an adversarially learned embedding mapping model that directly maps a short embedding to an enhanced embedding with increased discriminability. In particular, a Wasserstein GAN with a bunch of loss criteria are investigated. These loss functions have distinct optimization objectives and some of them are less favoured for the speaker verification research area. Different from most prior studies, our main objective in this study is to investigate the effectiveness of those loss criteria by conducting numerous ablation studies. Experiments on Voxceleb dataset showed that some criteria are beneficial to the verification performance while some have trivial effects. Lastly, a Wasserstein GAN with chosen loss criteria, without fine-tuning, achieves meaningful advancements over the baseline, with 4% relative improvements on EER and 7% on minDCF in the challenging scenario of short 2second utterances.

Index Terms— speaker embedding, speaker verification, generative adversarial network

1. INTRODUCTION

Text-independent Speaker Verification (SV) aims to automatically verify the identity of a speaker, given enrolled speaker record and some test speech signal (with no special constraint on phonetic content). The most important step in the SV pipeline is to map speech of arbitrary duration into speaker representation of fixed dimension. It's desirable for such a speaker representation to be compact, discriminative and robust to extrinsic and intrinsic variations.

Several types of speaker representations have been developed over the past decades. The well-known i-vector [3] has been the state-of-the-art speaker representation, usually associated with a simple cosine-scoring strategy or more powerful probability linear discriminant analysis (PLDA) [12, 4] as verifier. With the advent of deep neural networks (DNNs), a variety of DNN frameworks and loss functions have been developed to learn deep speaker representations, known as embeddings. By training these networks with either the cross-entropy loss, or some form of contrastive loss on large amount of data, the resulting embeddings are speaker-discriminative. Compared to the i-vector, those embeddings, such as x-vector[2] and GhostVLAD-aggregated embedding [18] (or G-vector for short), are promising, demonstrating competitive performance for long speeches and distinct advantage for short speeches. Furthermore, the recently developed G-vector further shows considerable gains over x-vector for noisy test conditions, which makes it more favorable for a practical SV system.

However, the performance of a SV system usually degrades in real scenarios, due to prevalent mismatches between development and test condition, such as channel, domain or duration mismatch [11, 5, 18]. For instance, it has been observed [5] that on NIST-SRE 2010 test set (female part), the performance of i-vector/PLDA system drops from 2.48% to 24.78% when the verification trial was shortened from full-duration to 5 seconds long.

Numerous research studies have been proposed to mitigate the short duration effect. An early family of researches aimed to modify different aspects of i-vector based SV system, e.g., feature extraction techniques, intermediate parameter estimation, speaker model generation, score normalization techniques, as summarized in [11]. Recently, more novel deep learning technologies are explored. For instance, insufficient phonetic information is compensated by a teacher-student learning framework [17] and scoring scheme is calibrated by transfer learning [13]. Another research strategy is to design duration robust speaker embeddings to dealing with utterances of arbitrary duration. By applying different neural network architectures and alternative loss functions, the discriminability of embeddings is further enhanced. For example, Inception Net with triplet loss is depolyed in [20], Inception-ResNet with joint softmax and center loss in [8] and ResCNN with novel speaker identity subspace loss in [14].

Generative Adversarial Networks (GANs) [6] are one of the most popular deep learning algorithm developed recently. GANs have the potential to generate realistic instances and provide a solution to problems that require a generative solution, most notably in various image-to-image translation tasks.

In this study, we aim to investigate the short duration is-

sue presented in a practical SV system. Contrary to the most techniques mentioned above, our proposed approach works directly on the speaker embeddings. In particular, given short and long embedding pairs extracted from same speaker and session, we propose to use adversarial learning of Wasserstein GAN to learn a new embedding with enhanced discriminability. To test our approach, G-vector is chosen as the embedding benchmark in our experiments due to its promising performance on short speeches. This put forward a challenge to our study than those prior studies which benchmarked with the i-vectors.

The remainder of this paper is organized as follows: Section 2 briefly introduces the related works of our methods. Section 3 details our proposed Wassertein GAN based approach. Section 4 presents experimental results and discussions. Finally, our conclusions are given in Section 5.

2. RELATED WORKS

2.1. Wasserstein-GAN

GANs [6] are deep generative models comprised of two networks, a generator and a discriminator. The discriminator D tries to learn the difference between real sample y and fake sample g generated from noise η, and the generator G tries to fool the discriminator. That is, the following minimax loss function is optimized through alternating optimization, until equilibrium is reached.

$$\min_G \max_D V_{GAN}(D,G) = E_y[logD(y)] + E_\eta[log(1-D(G(\eta)))] \quad (1)$$

However, training a GAN model is difficult due to well-known diminishing or exploding gradients issue. The issues has been addressed by Wasserstein GAN (WGAN) [1], where the discriminator is designed to find a good f_w and a new loss function is configured as measuring the Wasserstein distance:

$$W = \max_{f_w \in 1-Lipschitz} \{E_y[f_w(y)] - E_\eta[f_w(G(\eta))]\}$$

2.2. Deployments of GANs in SV

Motivated by the remarkable success in image-to-image translation, GANs have been actively deployed in SV research community, mainly to handle domain-mismatch issue, like transforming i-vectors [15] and x-vectors [19]. In contrast, there are few works to use GANs to handle the short duration issue. To authors' best knowledge, the only published work is to propose compensating the i-vectors via conditional GAN [7]. However, limited performance improvements were observed. The proposed system alone failed to outperform the baseline system, and only score-wise fusion based system showed better performance than the baseline.

In authors' opinion, training GAN is non-trivial, the reason behind such results might be the oversight on effects of loss functions of conditional GAN. As such, in this study, we investigate the problem and seek to reveal some guidelines on choosing beneficial loss functions to make the model perform better.

3. PROPOSED APPROACH

The architecture of our proposed approach is illustrated in Fig.1. Here x and y are D-dimensional G-vectors corresponding to short and long utterance embedding from same speaker session, z is speaker identity labels. With given x, y, z, the proposed system is trained to learn a D-dimensional embedding g, with the expectation that the g-based SV system can outperform the one based on x.

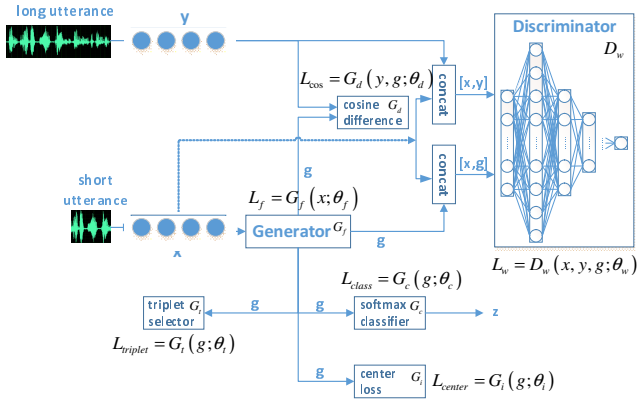

Fig. 1. Framework of our proposed system

Overall, the proposed architecture can be decomposed into four core components: embedding generator G_f, speaker label predictor G_c, distance calculator G_d and Wasserstine discriminator D_w. All components are jointly trained in order to generate enhanced embeddings with carefully handcraft optimization objects, as described as follows.

3.1. Proposed Discriminator-Related Loss Functions

As aforementioned, the primary task of the proposed approach is to learn embedding with enhanced discriminability. Let P denote the data distribution, we propose to achieve the task by mapping P_g from initial P_x to the target P_y by adversarial learning of WGAN. To this end, in the discriminative model, several loss criteria are investigated with different optimization objectives.

Following the conventional definition of min-max function, the loss function of WGAN is:

$$\min_{G_f} \max_{D_w} L_w(D_w, G_f) = E_y[D_w(y)] + E_x[D_w(G_f(x))] \quad (2)$$

Inspired by the idea of conditional GAN [10], in this study, we investigate a novel loss function by optimizing the

Wassertein distance between joint data distributions. That is, to control the data to be discriminated by concatenating short embedding x with the conventional discriminator input. The corresponding min-max function is updated as:

$$\min_{G_f}\max_{D_w} L_{cw}(D_w, G_f) = E_y[D_w(y;x)] + E_x[D_w(G_f(x);x))] \tag{3}$$

In addition, to seek more discriminability, the Fréchet Inception Distance (FID) [9], as a popular metric to calculate the distance between feature vectors of real and generated images, is also explored herein. Assuming P_y and P_g as normal distributions with means μ_y, μ_g and co-variance matrices C_y, C_g, FID loss can be calculated by:

$$L_{fid} = |\mu_y - \mu_g|^2 + tr\left(C_y + C_g - 2(C_yC_g)^{\frac{1}{2}}\right) \tag{4}$$

3.2. Proposed Generator-Related Loss Functions

In order to guide GAN training with the objective of feature discriminability, four loss criteria are investigated herein as extra training guides for the GAN training.

To verify the speaker label, the widely adopted multiclass cross-entropy (CE) loss is investigated with formulation of:

$$L_{class} = \frac{1}{N}\sum_{i=1}^{N} log\frac{e^{W_{z_i}^T g_i + b_{z_i}}}{\sum_{j=1}^{c} e^{W_j^T g_i + b_j}} \tag{5}$$

where N is the batch size, c is the number of classes. g_i denotes the i-th generated embedding sample and z_i is the corresponding label index. $W \in \Re^{D*c}$ and $b \in \Re^c$ denotes the weight matrix and bias in the project layer.

To explicitly penalize the class-related classification error, triplet loss is deployed as well, where a baseline (anchor) input is compared to a positive (truthy) input and a negative (falsy) input. Let Γ be the set of all possible embedding triplets $\gamma = (g_a, g_p, g_n)$ in the training set, the loss is defined as:

$$L_{triplet} = \sum_{\gamma\in\Gamma} max\left(\parallel g_a - g_p \parallel_2^2 - \parallel g_a - g_n \parallel_2^2 + \Psi, 0\right) \tag{6}$$

where g_a is an anchor input, g_p is a positive input from the same class and g_n is a negative input from a different class, $\Psi \in \Re^+$ is safety margin between positive and negative pairs.

Apart from the above, to minimize intra-class variation, center loss [16] is also adopted. It can be formulated as:

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

where c_{y_i} denotes the y_i^{th} class center of deep features, x_i denotes the ith deep feature belonging to the y_ith class and m is the size of mini-batch.

To better guide the training process, the similarity between enhanced embedding and its target is explicitly considered. It's measured by the cosine distance and evaluated as a dot product as follow:

$$L_{cos} = 1 - \bar{g}^*\bar{y} \tag{7}$$

where $\bar{g}$ and $\bar{y}$ are normalized version of embedding g and y, respectively.

In all, we propose to train the generator G_f with the total loss defined as:

$$L_G = L_w/L_{cw} + \alpha L_{class} + \beta L_{cos} + L_{center} + \epsilon L_{triplet} \tag{8}$$

and discriminator D_w with:

$$L_W = L_w/L_{cw} + \lambda L_{fid} \tag{9}$$

After the training of WGAN, the generative model G_f is retained. At the SV test stage, a short embedding x for any given test short utterance, can be easily mapped to its enhanced version (g) by directly applying the feed-forward model of G_f on the x.

4. EXPERIMENTS AND RESULTS

This section details our experimental setups and investigation results on the effectiveness of the above proposed loss criteria.

4.1. Experimental Setup

We use a subset of the Voxceleb2 to train our proposed system, where 1,057 speakers are chosen with total 164,716 utterances. Those utterances are randomly cut to 2 seconds as short utterance. Similarly, a subset of Voxceleb1 with 40 speakers is sampled and total 13,265 utterance pairs are used for testing.

The VGG-Restnet34s network is used to extract G-vectors as our baseline system. Regarding the GAN training, the learning rates for both G_f and D_w are 0.0001; Adam optimization is adopted; weight clipping is employed for G_w with threshold setting from -0.01 to 0.01 and batch size is set as 128.

4.2. Ablation Studies on Various Loss Functions

To verify the importance of proposed loss criteria, a bunch of ablation studies are conducted by choosing different combinations of them. The overall results are illustrated in Tab.1, where L_c, L_t denote L_{center} and $L_{triplet}$, respectively. Triplet a means that inputs are sampled from both y and g and b means from g only.

In our study, total 8 systems ($v1 - v8$), by combining different loss criteria with Watterstein GAN, are evaluated. Their corresponding detection error trade-off (DET) curves are plotted in Fig.2.

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	✓	✓	

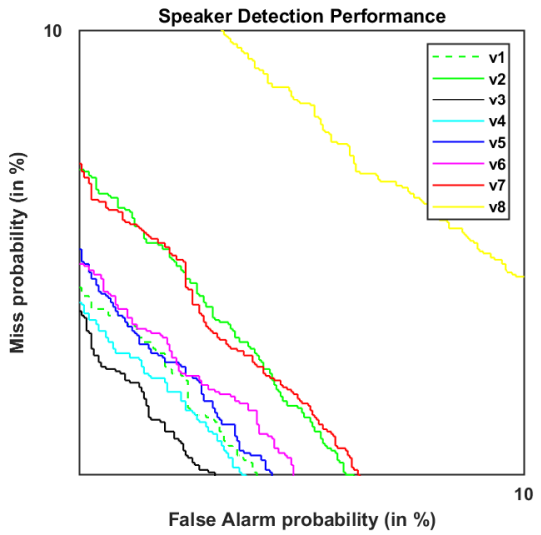

Fig. 2. DET performances for different systems

From the above experimental results, the following conclusions could be drawn:

- 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. $v8$).

The above findings are very interesting with a twofold outcome. Firstly, it demonstrates that additional training functions (e.g. traditional softmax, cosine loss and triplet loss) all have positive contribution to the performance, which verifies our earlier statement that extra training guides might be helpful for feature discriminability. Secondly, some less-favoured loss criteria to a typical SV system (e.g. FID loss and conditional WGAN loss) are surprisingly helpful, which are unusual findings and might be worthy of further investigation.

4.3. Comparison with the Baseline System

In the end, we make a performance comparison between our best system ($v3$) and the G-vector baseline system. Herein the comparison is measured in terms of equal error rate (EER) and minDCF. The results are reported in Tab.2.

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

From the table, we can see that our proposed system also has the merit for generalization and behave consistently for different short duration over the baseline system. In detail, for verification with 2 second enroll-test utterances, our proposed system shows 4.2% relative EER improvement and 7.2% relative minDCF improvement. For shorter utterances with duration of 1 second, it shows comparable EER (3.8%) improvement.

It's worth noting that due to time constraint, the FID loss function has not been added to our final system; besides, there is no any fine-tuning on hyper-parameters, loss weights $\alpha, \beta, \gamma, \lambda, \epsilon$ and triplet margin η. This means there are still a lot of room for improvements in our system.

5. CONCLUSIONS

In this paper, we have successfully applied WGAN to learn enhanced embedding for speaker verification application with short utterances. Our main contributions are twofold: proposed WGAN-based kernel system; and on top of it, validated the effectiveness of a bunch of loss criteria on the GAN training. Our final proposed system outperforms the baseline system for the challenging short speaker verification scenarios. In all, our experiments show both decent advancement and a potential direction where our further research goes forward.

6. REFERENCES

[1] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *ArXiv*, 2017. URL: https://arxiv.org/pdf/1701.07875.pdf.

[2] D. Povey D. Snyder, D. Garcia-Romero and S. Khudanpur. Deep neural network embeddings for text-

independent speaker verification. In *Interspeech*, page 999, 2017.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, May 2011.

[4] Daniel Garcia-Romero and Carol Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*, pages 249–252, 2011.

[5] Jahangir Alam Gautam Bhattacharya and Patrick Kenny. Deep speaker embeddings for short duration speaker verification. In *Interspeech, Stockholm*, pages 1517–1521, 2017.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014.

[7] Nakamasa Inoue Jiacen Zhang and Koichi Shinoda. I-vector transformation using conditional generative adversarial networks for short utterance speaker verification. In *Interspeech, Hyderabad*, pages 3613–3617, 2018.

[8] Na Li, Deyi Tuo, Dan Su, Zhifeng Li, and Dong Yu. Deep discriminative embeddings for duration robust speaker verification. In *Interspeech*, pages 2262–2266, 2018.

[9] Thomas Unterthiner Bernhard Nessler Martin Heusel, Hubert Ramsauer and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6626–6637, 2017.

[10] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv*, 2014. URL: https://arxiv.org/pdf/1411.1784.pdf.

[11] A. Poddar, M. Sahidullah, and G. Saha. Performance comparison of speaker recognition systems in presence of duration variability. In *2015 Annual IEEE India Conference (INDICON)*, pages 1–6, Dec 2015.

[12] S. J. D. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007.

[13] Lihong Wan Jun Zhang Qingyang Hong, Lin Li and Feng Tong. Transfer learning for speaker verification on short utterances. In *Interspeech*, pages 1848–1852, 2016.

[14] Xinyuan Cai Ruifang Ji and Bo Xu. An end-to-end text-independent speaker identification system on short utterances. In *Interspeech*, pages 3628–3632, 2018.

[15] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Chng, and Haizhou Li. Unsupervised domain adaptation via domain adversarial training for speaker recognition. In *ICASSP*, pages 4889–4893, 2018.

[16] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016.

[17] Jee weon Jung, Hee-Soo Heo, Hye jin Shim, and Ha-Jin Yu. Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings. *ArXiv*, 2018. URL: https://arxiv.org/pdf/1810.10884.pdf.

[18] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Utterance-level aggregation for speaker recognition in the wild. In *ICASSP*, pages 5791–5795, 2019.

[19] Man-Wai Mak Youzhi Tu and Jen-Tzung Chien. Variational domain adversarial learning for speaker verification. In *Interspeech*, pages 4315–4319, 2019.

[20] Chunlei Zhang and Kazuhito Koishida. End-to-end text-independent speaker verification with triplet loss on short utterances. In *Interspeech*, pages 1487–1491, 2017.