論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Generative Adversarial Network Based i-Vector Transformation for Short Utterance Speaker Verification
著者(和文)	ZHANG Jiacen, 井上 中順, 篠田 浩一
Authors(English)	Jiacen Zhang, Nakamasa Inoue, Koichi Shinoda
出典(和文)	2018年 秋季研究発表会 講演論文集, , , pp. 1345-1346
Citation(English)	ASJ 2018 Autumn Meeting, , , pp. 1345-1346
発行日 / Pub. date	2018, 8

Generative Adversarial Network Based i-Vector Transformation for Short Utterance Speaker Verification *

○ Jiacen Zhang, △ Nakamasa Inoue, Koichi Shinoda (Tokyo Institute of Technology)

1 Introduction

The speaker verification system extracts speaker characteristic information from a given utterance and then verify the speaker ID. In the state-of-the-art methods of speaker verification, i-vector [1] is used to represent speaker characteristics, and probabilistic linear discriminant analysis (PLDA) [2] is used as a verifier. While this system performs well on long utterances, the performance degrades drastically when only short utterances are available [3]. The main cause of this problem is the biased phonetic distribution of short utterances, which makes the estimated speaker features become statistically unreliable. This paper describes an i-vector transformation method using conditional GAN [4] for improving i-vector based short utterance speaker verification.

2 GAN for i-vector Transformation

Our target is to estimate a transformation function which can restore a reliable i-vector (extracted from a long utterance), from a short-utterance i-vector using conditional GAN. In the training stage, the generator G is optimized to generate a reliable i-vector using the one extracted from a short utterance, and the discriminator D is optimized to determine whether the given reliable i-vector is fake (generated by G) or real (extracted from a long utterance). In test, G is used as the transformation function for an i-vector extracted from a short utterance in the test set.

We use a special GAN structure, Wasserstein GAN (WGAN) [5]. Denoting x as an unreliable i-vector, y as a reliable i-vector and z as random noise, the min-max function is represented as:

$$\min_G \max_D V(D,G) = E_{x,y}D(y|x) - E_{x,z}D(G(z|x)), \tag{1}$$

and the objective function for G and D are,

$$\min \mathrm{G} = -E_{x,z}D(G(z|x)), \tag{2}$$

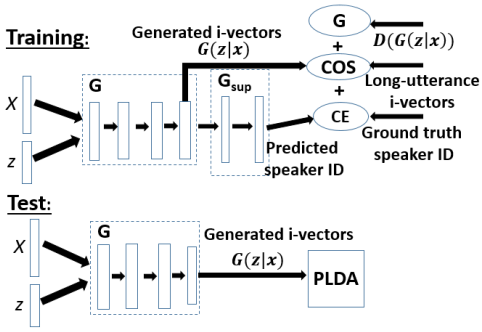

Fig. 1　Training and test stage of G.

$$\max \mathrm{D} = E_{x,y}D(y|x) - E_{x,z}D\left(G\left(z|x\right)\right). \tag{3}$$

Regarding the training data for GAN, i-vectors extracted from short and long utterances are required. While only long utterances are present in the training dataset, we obtained short utterances by segmenting a long utterance into short utterances, so we can obtain an i-vector pair consisting two i-vectors from the same speaker and session, but one is from a short utterance and the other is from a long utterance.

To better guide the training of GAN for our task and make the best use of the training data, two additional learning tasks are added to the GAN framework. The first one is to minimize the cosine distance between the generated i-vector and the target:

$$\min \mathrm{COS} = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\left(1 - \frac{G(z|x_{ij})\cdot y_i}{\|G(z|x_{ij})\|\ \|y_i\|}\right)\right], \tag{4}$$

where m is the number of long utterances in the training set, y_i is the i-vector extracted from the i-th long utterance in the training set, n_i is the number of short utterances got from the i-th long utterance, x_{ij} is the i-vector extracted from the j-th segment of the i-th long utterance and z is random noise.

Obviously, improving the speaker discriminative ability of generated i-vectors can enlarge the inter-speaker differences among i-vectors and improve the verification performance in the PLDA scoring stage. As shown in Figure 1, in the training stage, a supplementary section, G_{sup}, is concatenated after the

*短い発話における話者照合のための敵対的生成ネットワークを用いた i-vector 変換. 張佳岑、井上中順、篠田浩一 (東京工業大学)

Table 1 The speaker verification results in terms of EER (%) on the SRE08 "short2-10sec" and "10sec-10sec" condition 6 male trail list.

System	short2-10sec	10sec-10sec
a) Baseline	7.28	11.97
b) D-WCGAN	9.45	15.42
c) a + b	**6.89**	**10.75**

generator G, which takes the generated i-vector as an input and predicts its speaker label. We minimize cross entropy between the prediction result and the ground truth:

$$\min \mathrm{CE} = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n_i}\sum_{j=1}^{n_i} l_{ij}^{k}\left(\log o_{ij}^{k}\right)\right], \quad (5)$$

where l_{ij}^{k} is the empirical probability observed in the ground truth that the target i-vector belongs to the k-th class, and o_{ij}^{k} is the predicted probability that the generated i-vector belongs to the k-th class. In summary, for training G, our goal is to minimize

$$a\ \mathrm{G} + b\ \mathrm{COS} + c\ \mathrm{CE}, \quad (6)$$

where a, b, c are weight parameters for these three targets, respectively. After training, as shown in Figure 1, only G is used to generate a reliable i-vector, which is fed into a PLDA model for the next scoring step.

3 Evaluation

We evaluated the performance of our method using the equal error rate (EER) calculated on the "short2-10sec" and "10sec-10sec" condition 6 male trials of the NIST SRE 2008 [6]. The i-vector and PLDA are trained with SRE08's development data, which contains the NIST SRE2004-2006 data, Switchboard, and Fisher corpus. This dataset as a whole consistes 34,925 utterances from 7,275 male speakers. The baseline is this system without any compensation. The training data of the GAN is a subset of SRE08's development set mentioned above and SRE08's training set, which contains 1,986 male speakers in total. To make the short and long utterance pairs mentioned in Section 4, we used a sliding window of 20s long and 10s shift to cut one long utterance into short utterances. Finally, we got 331,675 i-vector pairs for GAN training.

Table 1 shows the results of the experiments. Although our method (the discriminative Wasserstein Conditional GAN, D-WCGAN) alone did not outperform the baseline system, it achieved better results when the score-wise fusion was done with the baseline method. The results was achieved when the score weight ratio of the baseline system and our method is 7:3. These results showed that our proposed method can help make i-vectors more reliable in most cases. However, in the current stage, the amount of training data for the GAN is not enough, even smaller than the amount of PLDA's training data. If we have more training data for the GAN, the performance of the proposed methods may become much better.

4 Conclusions

This paper has proposed a GAN-based speaker feature restoration method for speaker verification using short utterances. The generator is trained to transform an unreliable i-vector extracted from a short utterance to a reliable i-vector which can be extracted from a long utterance. Speaker labels are also used in the training of GAN to improve the speaker discriminative ability of generated i-vectors. The evaluation results on NIST SRE 2008 task show that our proposed method improved the performance of i-vector and PLDA based short utterance speaker verfication.

Acknowledgement This work was supported by JSPS KAKENHI 16H02845 and by JST CREST Grant Number JPMJCR1687, Japan.

References

[1] Dehak *et al.*, IEEE Transactions on Audio, Speech, and Language Processing, 19 (4), 788-798, 2011.
[2] Kenny, Proc. of Odyssey Speaker and Language Recognition Workshop, 28-33, 2010.
[3] Kanagasundaram *et al.*, Proc. of Interspeech, 2341-2344, 2011.
[4] Mirza, *et al.*, arXiv preprint arXiv:1411.1784, 2014.
[5] Arjovsky, *et al.*, Proc. of ICML, 214-223 , 2017.
[6] https://www.itl.nist.gov/iad/mig/tests/spk/2008/index.html.