Data Augmentation using Variational Autoencoder for Embedding based Speaker Verification

Zhanghao Wu, **Shuai Wang**, Yanmin Qian, Kai Yu

Speech Lab, Shanghai Jiao Tong University, China

September 2019

- Speaker embeddings are now the main approach for speaker identity modelling
- SV systems still suffer from performance degradation due to the complex environment in real applications
- How to improve the noise-robustness of the SV systems?
  - Data augmentation is proved to be simple but effective
  - A robust PLDA back-end is also helpful

Overview

Basic Idea

- Use **Variational Auto-Encoder** to generate more diverse speaker embeddings
- Train a more robust PLDA with the augmented speaker embeddings
- Why at embedding level?
  - The final representation used for scoring
  - Get rid of the complexity of tying different frames
  - Simple yet effective

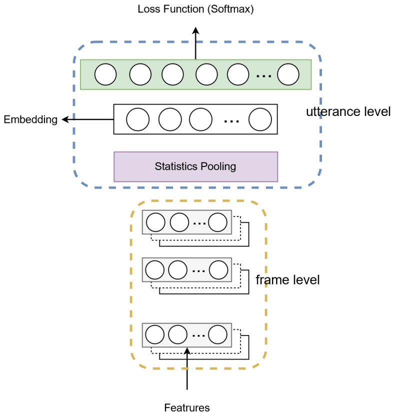
i-vector:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{x} + \epsilon,$$

PLDA

$$\mathbf{x}_j^{(s)} \sim \mathcal{N}(\mathbf{y}^{(s)}, \mathbf{W}^{-1})$$
$$\mathbf{y}^{(s)} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{B}^{-1})$$

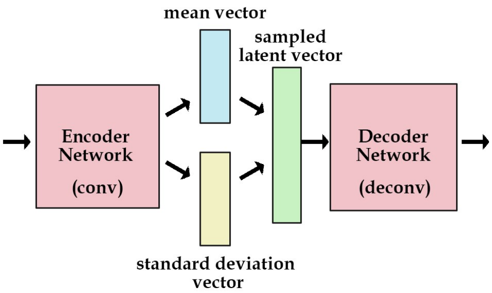
x-vector:



Related Work

Traditional Data Augmentation Method

1. Manually add noise to the raw audios
2. Generate more features from the augmented audios, train a speaker embedding extractor in the normal way
3. Extract the embeddings from augmented audios, train a noise-robust PLDA



- Widely used generative model
- Generate new samples with the decoder network

Can we use it to generate more diverse speaker embeddings?

[1] Kingma *et al.*, Auto-Encoding Variational Bayes

Conditional VAE[2]

CVAE for speaker embedding generation

- The generation process should preserve speaker identity
- Use conditional VAE, which conditions on speaker identity
- The target for the CVAE model is to maximize the likelihood of generated noise speaker embeddings conditioned on the clean embeddings.
- By sampling from normal distribution, we can generate noisy speaker embeddings based on a given clean speaker embedding with the CVAE model(the decoder part).

[2] Sohn *et al.*, Learning Structured Output Representation using Deep Conditional Generative Models

Lower bound of log-likelihood in VAE:

$$\begin{aligned}\log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\log q_\phi(\mathbf{z}|\mathbf{x}) + \log p_\theta(\mathbf{x},\mathbf{z})] \\ &= -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\end{aligned}$$

Introducing conditions:

$$\log p_\theta(\mathbf{x}|\mathbf{c}) \geq -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})||p_\theta(\mathbf{z}|\mathbf{c})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{c})]$$

$\mathbf{z}$: latent variable, $\mathbf{x}$: data from the dataset, $\mathbf{c}$: condition.

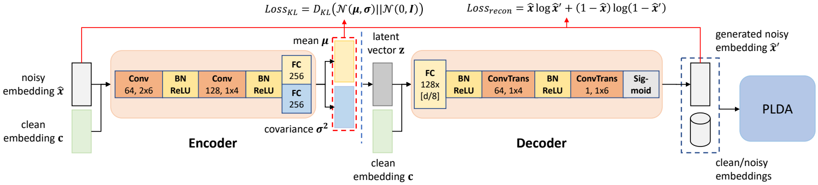


Figure: Framework and detailed neural network configuration of the proposed CVAE based data augmentation.

$$\mathcal{L}_{KL} = D_{KL}(q_\phi(\mathbf{z}|\hat{\mathbf{x}},\mathbf{y})||\mathcal{N}(0,\mathbf{I}))$$

$$\mathcal{L}_{recon} = \mathrm{BCE}(\hat{\mathbf{x}}_u^{(s)}, \hat{\mathbf{x}}_u'^{(s)})$$

$$\mathcal{L}_{total} = \mathcal{L}_{KL} + \mathcal{L}_{recon}$$

s: s-th speaker, u: u-th utterance, y: clean speaker embedding.

Dataset

Training data:
SWBD + SRE

Evaluation data:
SRE16 evaluation set

Training Settings:
All speaker embedding systems are trained on both training data. The PLDA and CVAE are only trained on SRE.

CVAE model

- Condition on the clean speaker embedding and trained on the manually augmented data.
- Each clean embedding corresponds to 4 noisy embeddings extracted from the manually augmented audios(Reverb, MUSAN noise, music, and speech).

Table: Performance comparison for *i*-vector/PLDA SV system using different data augmentation methods. The amount of augmented data for different methods are comparable.

	Data Augmentation	SRE16 Tagalog		SRE16 Cantonese	
		EER (%)	minDCF	EER (%)	minDCF
PLDA	none	18.13	0.7068	9.82	0.3951
+Adaptation		17.84	0.6338	8.82	0.3591
PLDA	manual	17.63	0.6961	9.42	0.3827
+Adaptation		16.94	0.6105	8.30	0.3411
PLDA	VAE	17.45	0.7185	10.14	0.4088
+Adaptation		15.83	0.5981	8.32	0.3461
PLDA	VAE & manual	17.20	0.7106	9.62	0.3940
+Adaptation		**15.54**	**0.5897**	**7.84**	**0.3331**

Table: Performance comparison of different data augmentation methods for *x*-vector/PLDA based SV system.

	Data Augmentation	SRE16 Tagalog		SRE16 Cantonese	
		EER (%)	minDCF	EER (%)	minDCF
PLDA	none	16.63	0.7121	7.57	0.3451
+Adaptation		14.10	0.5420	5.77	0.2523
PLDA	manual	16.16	0.7248	7.45	0.3368
+Adaptation		12.79	0.5144	5.26	0.2357
PLDA	GAN	16.54	0.7004	7.09	0.3363
+Adaptation		12.42	0.5196	4.66	0.2379
PLDA	GAN & manual	16.59	0.7182	6.85	0.3256
+Adaptation		**11.68**	0.4886	4.43	0.2160
PLDA	VAE	16.44	0.7150	6.705	0.3187
+Adaptation		12.04	0.4844	4.29	0.2051
PLDA	VAE & manual	16.13	0.7114	6.60	0.3082
+Adaptation		11.86	**0.4799**	**4.20**	**0.2032**

Results: Detection Error Trade-off

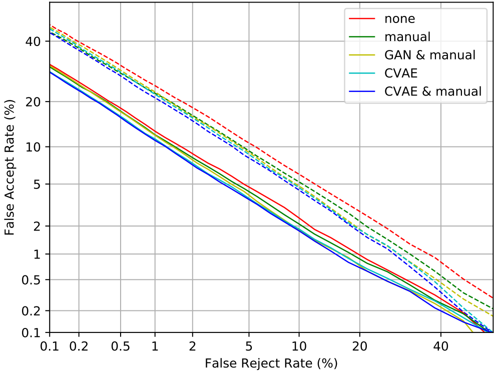


Figure: DET on Cantonese for *x*-vector based system. The dotted and concrete lines represent the non-adapted and adapted PLDA systems respectively.

Conclusions

- We proposed to use conditional variational autoencoder for data augmentation in the speaker verification task.
- Different from most data augmentation methods which are operated on the input audios, we directly augment the speaker embeddings and aim to train a more robust PLDA
- Our proposed model achieves promising results for both i-vector and x-vector framework.