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
Overview

Background

- 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
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
**Related Work**

**Embedding based Speaker Verification**

**i-vector:**

\[ M = m + Tx + \epsilon, \]

**PLDA**

\[ x_j^{(s)} \sim \mathcal{N}(y^{(s)}, W^{-1}) \]

\[ y^{(s)} \sim \mathcal{N}(\mu, B^{-1}) \]
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
Related Work

Variational Autoencoder

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

Can we use it to generate more diverse speaker embeddings?

\[^1\text{Kingma et al., Auto-Encoding Variational Bayes}\]
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).

---

Sohn et al., Learning Structured Output Representation using Deep Conditional Generative Models
Lower bound of log-likelihood in VAE:

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

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

Introducing conditions:

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

z: latent variable, x: data from the dataset, c: condition.
Data Augmentation with CVAE

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

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

\[
\mathcal{L}_{recon} = \text{BCE}(\hat{x}_u^{(s)}, \hat{x}_u'^{(s)})
\]

\[
\mathcal{L}_{total} = \mathcal{L}_{KL} + \mathcal{L}_{recon}
\]

\( s: \) \( s \)-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.

<table>
<thead>
<tr>
<th>Data Augmentation</th>
<th>SRE16 Tagalog</th>
<th>SRE16 Cantonese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER (%)</td>
<td>minDCF</td>
</tr>
<tr>
<td>PLDA none</td>
<td>18.13</td>
<td>0.7068</td>
</tr>
<tr>
<td>+Adaptation</td>
<td>17.84</td>
<td>0.6338</td>
</tr>
<tr>
<td>PLDA manual</td>
<td>17.63</td>
<td>0.6961</td>
</tr>
<tr>
<td>+Adaptation</td>
<td>16.94</td>
<td>0.6105</td>
</tr>
<tr>
<td>PLDA VAE</td>
<td>17.45</td>
<td>0.7185</td>
</tr>
<tr>
<td>+Adaptation</td>
<td>15.83</td>
<td>0.5981</td>
</tr>
<tr>
<td>PLDA VAE &amp; manual</td>
<td>17.20</td>
<td>0.7106</td>
</tr>
<tr>
<td>+Adaptation</td>
<td><strong>15.54</strong></td>
<td><strong>0.5897</strong></td>
</tr>
</tbody>
</table>
## Experiments

### Results: Augmenting x-vector/PLDA SV system

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

<table>
<thead>
<tr>
<th>Data Augmentation</th>
<th>SRE16 Tagalog</th>
<th>SRE16 Cantonese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER (%)</td>
<td>minDCF</td>
</tr>
<tr>
<td>PLDA none</td>
<td>16.63</td>
<td>0.7121</td>
</tr>
<tr>
<td>+Adaptation</td>
<td>14.10</td>
<td>0.5420</td>
</tr>
<tr>
<td>PLDA manual</td>
<td>16.16</td>
<td>0.7248</td>
</tr>
<tr>
<td>+Adaptation</td>
<td>12.79</td>
<td>0.5144</td>
</tr>
<tr>
<td>PLDA GAN</td>
<td>16.54</td>
<td>0.7004</td>
</tr>
<tr>
<td>+Adaptation</td>
<td>12.42</td>
<td>0.5196</td>
</tr>
<tr>
<td>PLDA GAN &amp; manual</td>
<td>16.59</td>
<td>0.7182</td>
</tr>
<tr>
<td>+Adaptation</td>
<td><strong>11.68</strong></td>
<td><strong>0.4886</strong></td>
</tr>
<tr>
<td>PLDA VAE</td>
<td>16.44</td>
<td>0.7150</td>
</tr>
<tr>
<td>+Adaptation</td>
<td>12.04</td>
<td>0.4844</td>
</tr>
<tr>
<td>PLDA VAE &amp; manual</td>
<td>16.13</td>
<td>0.7114</td>
</tr>
<tr>
<td>+Adaptation</td>
<td><strong>11.86</strong></td>
<td><strong>0.4799</strong></td>
</tr>
</tbody>
</table>
Experiments

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.