Short-time Spectral Aggregation for Speaker Embedding

Youzhi TU and Man-Wai MAK

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR of China

ICASSP’21
6-11 June 2021
Contents

1. Speaker embedding networks
2. Conventional pooling methods
3. Short-time spectral pooling (STSP)
4. Experimental setup
5. Results
6. Conclusions
Speaker Embedding Networks

- X-vector extractor is a popular baseline
  - Frame-level layers: Time-delay neural networks (TDNNs), ResNets, DenseNets, Res2Nets, etc.
  - Pooling layer: Aggregate frame-level information
  - Utterance-level layers: Fully-connected (FC) layers

Each layer except the pooling layer is followed by a batch normalization layer and an ReLU layer.
Pooling Methods

- Input: Temporal feature maps $\mathbf{X} \in \mathbb{R}^{C \times T}$ at the output of the last frame-level layer, $C$ and $T$ are the number of channels and frames, respectively.
- Output: Aggregated representation $\mathbf{z}$ at utterance-level.
- Statistics pooling: $\mathbf{z}$ is the concatenation of channel-wise mean and standard deviation (std).

![Diagram of pooling process](image)
**Pooling Methods**

- **Attentive pooling (AP):** Attend to discriminative frames
  - The aggregated representation $z$ is the concatenation of weighted channel-wise mean and standard deviation
  - The attention weight vector (for a single head) $w \in \mathbb{R}^{1 \times T}$ is learned from an attention network and applied to the features of each channel
  - For multi-head attentive pooling, $z$ is the concatenation of the aggregated representations corresponding to different heads
Motivation

• Limitation of statistics pooling
  • Using means and standard deviations is not enough to preserve sufficient speaker information for statistics pooling
  • From a Fourier perspective, statistics pooling only exploits the information in the 0-th frequency component (DC component) in the spectral domain

• Solution
  • Extract multiple spectral components of the spectral representation (besides the DC component) as aggregated embeddings
Short-time Spectral Pooling (STSP)

Perform STFT along the temporal axis for each channel

Average $|X_c(m, k)|$ along the temporal axis

Average $|X_c(m, k)|^2$ along the temporal axis

Concatenate $\hat{X}_c(0)$ and the lowest $R$ components of $P_c(k)$ from all channels

Aggregated statistics $z$
Relation to Statistics Pooling

• Short-time Fourier transform (STFT) of the $c$-th channel feature $x_c = \{x_c(n)\}_{n=0}^{N-1}$ ($N$ is the number of frames)

$$X_c(m, k) = \sum_{n=0}^{N-1} x_c(n) \omega(n - mS)e^{-j\frac{2\pi}{L}kn}, \quad k = [0, L-1]$$

$\omega(\cdot)$: window function, $L$: STFT length, $S$: step size of the sliding window
$m$: index of windowed segments, $k$: index of spectral components

• When we use $\omega(n) = 1$ (rectangular window) and $S = L = 1$ (the step size and STFT length are both 1), we have

$$\hat{X}_c(0) = 1/M \sum_{m=0}^{M-1} X_c(m, 0) = 1/N \sum_{n=0}^{N-1} x_c(n) \triangleq \text{mean}(x_c),$$

$$P_c(0) = 1/M \sum_{m=0}^{M-1}|X_c(m, 0)|^2 = 1/N \sum_{n=0}^{N-1}[x_c(n)]^2 \triangleq \text{var}(x_c) + [\text{mean}(x_c)]^2.$$  

• Under above conditions, using means and stds for statistics pooling is an analogy to using the DC components $\hat{X}_c(0)$ and $P_c(0)$ for STSP

• Because STSP uses more frequency components of $P_c(k)$ ($k > 0$) for aggregation, it can preserve more information than statistics pooling
Experiments

• Compare statistics pooling, attentive pooling and STSP on VoxCeleb1-test, VOiCES19-dev and VOiCES19-eval

• Speaker embedding network training
  • 40-dimensional filter bank features
  • VoxCeleb1&2-dev for VOiCES19 (2,105,949 utterances from 7,185 speakers) and VoxCeleb1-dev for VoxCeleb1 (2,092,009 utterances from 5,984 speakers)
  • Baseline: Standard x-vector network
  • Attention network: FC (500) + ReLU + FC (H), H is the number of heads
  • STSP: Rectangular window function, STFT length and window step size were 16

https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2
Experiments

• PLDA training
  • VoxCeleb1: Clean VoxCeleb1-dev (1,240,651 utterances)
  • VOiCES19: Concatenated speech with the same video session augmented with reverberation and noise (334,776 utterances)
  • Pre-processing: Center + LDA (200 for Voxceleb1 and 150 for VOiCES19) + whitening + length normalization

• Score normalization (only for VOiCES19)
  • Cohort: Longest two utterances of each speaker in the PLDA training data
Results on Voxceleb1-test

$H$: Number of heads in attentive pooling

$R$: Number of spectral components of $P_c(k)$ in STSP

<table>
<thead>
<tr>
<th></th>
<th>Stats pooling</th>
<th>AP ($H=1$)</th>
<th>AP ($H=2$)</th>
<th>AP ($H=3$)</th>
<th>AP ($H=4$)</th>
<th>STSP ($R=1$)</th>
<th>STSP ($R=2$)</th>
<th>STSP ($R=3$)</th>
<th>STSP ($R=4$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EER</strong></td>
<td>2.13</td>
<td>2.05</td>
<td><strong>1.96</strong></td>
<td>1.99</td>
<td>2.01</td>
<td>2.17</td>
<td>1.91</td>
<td><strong>1.82</strong></td>
<td>1.93</td>
</tr>
<tr>
<td><strong>minDCF</strong></td>
<td>0.227</td>
<td>0.221</td>
<td><strong>0.207</strong></td>
<td>0.218</td>
<td>0.232</td>
<td>0.221</td>
<td><strong>0.199</strong></td>
<td>0.210</td>
<td>0.220</td>
</tr>
</tbody>
</table>
Results on VOiCES19-dev

$H$: Number of heads in attentive pooling  
$R$: Number of spectral components of $P_c(k)$ in STSP

<table>
<thead>
<tr>
<th></th>
<th>Stats pooling</th>
<th>AP ($H=1$)</th>
<th>AP ($H=2$)</th>
<th>AP ($H=3$)</th>
<th>AP ($H=4$)</th>
<th>STSP ($R=1$)</th>
<th>STSP ($R=2$)</th>
<th>STSP ($R=3$)</th>
<th>STSP ($R=4$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER</td>
<td>2.32</td>
<td>2.40</td>
<td>2.10</td>
<td>2.09</td>
<td>2.12</td>
<td>2.25</td>
<td>2.05</td>
<td>2.16</td>
<td>2.08</td>
</tr>
<tr>
<td>minDCF</td>
<td>0.273</td>
<td>0.291</td>
<td><strong>0.270</strong></td>
<td><strong>0.270</strong></td>
<td>0.292</td>
<td>0.280</td>
<td>0.283</td>
<td><strong>0.266</strong></td>
<td>0.275</td>
</tr>
</tbody>
</table>

![Graph showing EER and minDCF]
Results on VOiCES19-eval

\( H \): Number of heads in attentive pooling
\( R \): Number of spectral components of \( P_c(k) \) in STSP

<table>
<thead>
<tr>
<th></th>
<th>Stats pooling</th>
<th>AP ((H=1))</th>
<th>AP ((H=2))</th>
<th>AP ((H=3))</th>
<th>AP ((H=4))</th>
<th>STSP ((R=1))</th>
<th>STSP ((R=2))</th>
<th>STSP ((R=3))</th>
<th>STSP ((R=4))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EER</strong></td>
<td>6.19</td>
<td>6.02</td>
<td><strong>5.72</strong></td>
<td>5.79</td>
<td>5.92</td>
<td>6.20</td>
<td><strong>5.67</strong></td>
<td>5.76</td>
<td>5.84</td>
</tr>
<tr>
<td><strong>minDCF</strong></td>
<td>0.467</td>
<td><strong>0.465</strong></td>
<td>0.468</td>
<td>0.484</td>
<td>0.514</td>
<td><strong>0.469</strong></td>
<td>0.478</td>
<td>0.473</td>
<td>0.488</td>
</tr>
</tbody>
</table>

**Graphs:**
- EER vs Stats pooling, AP, STSP for different values of \( H \) and \( R \)
- minDCF vs Stats pooling, AP, STSP for different values of \( H \) and \( R \)
Conclusions

• Proposed a new pooling method for speaker embedding from a Fourier perspective

• STSP is able to aggregate the information in higher frequency components (besides the DC component), making it preserve more speaker information than statistics pooling

• Generally, STSP outperforms attentive pooling and statistics pooling on Voxceleb1 and VOiCES19
References


Thank you!