Speaker2Vec: Unsupervised Learning and Adaptation of a Speaker Manifold using Deep Neural Networks with an Evaluation on Speaker Segmentation

Arindam Jati and Panayiotis Georgiou, SCUBA Lab.

Speaker characteristics learning

Learning speaker-specific characteristics:
- Various applications in Speaker recognition, Segmentation and Diarization, and Automatic speech recognition
- MFCC or PLP as state-of-the-art features, but they contain many other information than speaker characteristics

Deep Neural Networks (DNN) based approaches:
- Bottleneck layer as “speaker embedding”
- drawbacks of supervised DNN based methods:
  - Need lot of labeled data, overfitting issues
  - Lack of robustness if test condition is different

Speaker2Vec

Goal:
- Learn speaker-specific characteristics

Completely unsupervised method, learn from a dataset that even contain multi-speaker audio streams

Benefits:
- Doesn’t require any labeled training data, highly scalable
- Unsupervised domain adaptation
- Voice activity detection (VAD) not needed for training

Idea: Neural Predictive Coding (NPC)

Methodology

- Predict one speech segment of a speaker from another speech segment of the same speaker
- Consecutive windows are from the same speaker if we have single-speaker audio streams
- Problem: How to tackle multi-speaker audio streams?
  Solution: Short-term active-speaker stationarity hypothesis

- Very unlikely that the speaker turns will occur very frequently (for example, every 1 second)
- It is highly probable that most of the pairs will have both the windows from one speaker

Training framework

Evaluation

- Evaluation: Speaker segmentation
- Using simple KL divergence based speaker segmentation method
- Just replace MFCC by the embeddings
- Metric: Equal Error Rate (EER) reported as (MDR, FAR)

Unsupervised Domain Adaptation

Find speaker change points using the trained model, do over-segmentation
Get speaker homogeneous regions, reject segments shorter than 2 training window
Retrain DNN using these regions

Experiments

Training datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size (hours)</th>
<th>Model</th>
<th>Parameters in DNN model</th>
</tr>
</thead>
<tbody>
<tr>
<td>TED-LIUM (TED)</td>
<td>118</td>
<td>Small</td>
<td>24M</td>
</tr>
<tr>
<td>YouTube (YT)</td>
<td>911</td>
<td>Small</td>
<td>24M</td>
</tr>
<tr>
<td>YouTubeLarge (YTL)</td>
<td>1953</td>
<td>Large</td>
<td>52M</td>
</tr>
</tbody>
</table>

DNN architectures (# hidden units in each layer)

- Small network for TED-LIUM and YouTube datasets:
  - [4000 → 2000 → 40 → 2000 → 4000]
- Large network for YouTubeLarge dataset:
  - [4000 → 6000 → 2000 → 40 → 2000 → 6000 → 4000]

Results: Comparison with baselines

Evaluatio

datasets

Different distance metrics with MFCC features

Speaker2Vec model

Adapted Speaker2Vec model

TED-LIUM

test

Bi with MFCC3 [47.5, 65.4] 47.5 65.4
GD with MFCC3 [60.0, 61.5] 60.0 61.5
KL2 with MFCC3 [62.5, 63.9] 62.5 63.9
KL with MFCC4 [75.8, 76.09] 75.8 76.09

Ted Y 43.2 44.0 44.3
Ted Y 43.0 44.0 44.4
Ted Y 33.9 44.0 44.4

CoupleTherapy

RT-06 bol. [51.7, 50.8] 51.7 50.8
RT-06 bol. [51.7, 50.8] 51.7 50.8
RT-06 bol. [51.7, 50.8] 51.7 50.8

Mean improvement w.r.t. best baseline

9.92 6.62 5.73 11.72 11.14 9.44

Evaluation: Speaker segmentation

Using simple KL divergence based speaker segmentation method
Just replace MFCC by the embeddings
Metric: Equal Error Rate (EER) reported as (MDR, FAR)

Conclusions:

- Completely unsupervised method to learn speaker-specific characteristics from unlabeled speech
- Outperformed MFCC based baselines and state-of-the-art methods in speaker segmentation
- Unsupervised domain adaptation technique helped improving the performance

Future directions:

- Applying the embeddings in speaker diarization and speaker recognition tasks
- Employing the method in a different domain. Some motivating results in Behavioral Signal Processing (BSP) domain with
  - acoustic (Li et. al., 2017) and
  - lexical (Tseng et. al., 2017) features

Summary

Experiments

Training framework

<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>Duration (mins)</th>
<th># change points</th>
<th>Unsupervised?</th>
<th>F1 score</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes/Yes/No/No</td>
<td>0.73/0.78/0.74/0.79</td>
<td>0.82/0.83/0.86/0.85</td>
</tr>
<tr>
<td>Speaker2Vec model</td>
<td>TED</td>
<td>6.42</td>
<td>60</td>
<td>No</td>
<td>0.73</td>
<td>15.6/28.2</td>
</tr>
<tr>
<td>Adapted Speaker2Vec model</td>
<td>TED</td>
<td>6.42</td>
<td>60</td>
<td>No</td>
<td>0.73</td>
<td>15.6/28.2</td>
</tr>
</tbody>
</table>

http://scuba.usc.edu/