Speaker Verification

Hung-yi Lee

Some slides are from 袁培傑
One slide for this course
Related Tasks

**Emotion Recognition**

- Model \(\rightarrow\) happy? sad?

**Sound Event Detection**

- Model \(\rightarrow\) Door Opening?

**Autism Recognition**

- Model \(\rightarrow\) autism?

**Keyword Spotting**

- Model \(\rightarrow\) Keyword Exist?

We only focus on **speaker verification** today.
Outline

Task Introduction

Speaker Embedding

End-to-end
Task Introduction

• Speaker Recognition / Identification
  • 語者識別
  • 一段語音是誰所說的

• Speaker Verification
  • 語者驗證
  • 兩段語音是否為同一人所說

• Speaker Diarization
  • 語者分段標記
  • 在一段語音中，誰在何時說話
Task Introduction

• Speaker Recognition / Identification
  • 語者識別
  • 一段語音是誰所說的

A multi-class classification problem
Task Introduction

• Speaker Recognition / Identification
  • 語者識別
  • 一段語音是誰所說的

• Speaker Verification
  • 語者驗證
  • 兩段語音是否為同一人所說
Speaker Verification

Enrollment

Evaluation

Model

> threshold?  
Same

< threshold?  
Different

Application: 銀行客服
## Equal Error Rate (EER)

<table>
<thead>
<tr>
<th>Threshold</th>
<th>True Positive (TP)</th>
<th>False Positive (FP)</th>
<th>False Negative (FN)</th>
<th>True Negative (TN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>0.8</td>
<td>30</td>
<td>23</td>
<td>70</td>
<td>77</td>
</tr>
<tr>
<td>0.6</td>
<td>50</td>
<td>34</td>
<td>50</td>
<td>67</td>
</tr>
<tr>
<td>0.4</td>
<td>78</td>
<td>52</td>
<td>22</td>
<td>48</td>
</tr>
<tr>
<td>0.2</td>
<td>84</td>
<td>76</td>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Equal Error Rate (EER)**

EER = 0.40

**False Negative Rate**

同一語者被判斷成不同語者

**False Positive Rate**

不同語者被判斷成同一語者

![Graph showing Equal Error Rate (EER)](image)
Task Introduction

• Speaker Recognition / Identification
  • 語者識別
  • 一段語音是誰所說的

• Speaker Verification
  • 語者驗證
  • 兩段語音是否為同一人所說

• Speaker Diarization
  • 語者分段標記
  • 在一段語音中，誰在何時說話

**diarize**: to write down your future arrangements, meetings, etc. in a diary
Speaker Diarization

Record of meeting, record of telephone conversion, etc.

Step 1: Segmentation

Step 2: Clustering

The number of speakers can be known or unknown.
Task Introduction

• Speaker Recognition / Identification
  • 語者識別
  • 一段語音是誰所說的

• Speaker Verification
  • 語者驗證
  • 兩段語音是否為同一人所說

• Speaker Diarization
  • 語者分段標記
  • 在一段語音中，誰在何時說話
Some Method

Calculate Similarity

Model

Same

Different

> threshold?

< threshold?

Scalar

Speaker Embedding

Calculate Similarity

Scalar
Framework

The speakers in stages 2 and 3 are not seen in stage 1.

Stage 1: Development

Stage 2: Enrollment

Stage 3: Evaluation

Having multiple enrollment utterances?

Calculate Similarity

Scalar
Metric-based meta learning

- https://youtu.be/yyKaACh_j3M
**Framework**

The speakers in stages 2 and 3 are not seen in stage 1.

**Stage 1: Development**

- Google’s Dataset (private) [Wan, et al., ICASSP’18]
  - 36M utterances, 18,000 speakers
- VoxCeleb [Nagrani, et al., INTERSPEECH’17]
  - 0.15M utterances, 1,251 speakers
- VoxCeleb2 [Chung, et al., INTERSPEECH’18]
  - 1.12M utterances, 6,112 speakers
i-vector

“i” means “identity”
d-vector

Training Speaker Recognition Model

Which Speaker?

output layer

DNN

audio segment

whole utterance
d-vector

[Variani, et al., ICASSP’14]

d-vector and i-vector are only comparable

average

DNN
DNN
DNN
DNN
DNN
DNN
Which Speaker?

\[ x\text{-vector} \]

[Snyder, et al., ICASSP’18]

**x-vector**

statistical pooling

\[
\text{mean} \rightarrow \text{DNN} \rightarrow \text{output layer} \rightarrow \text{Which Speaker?}
\]

whole utterance
Attention Mechanism

[Chowdhury, et al., ICASSP’18]

NetVLAD

[Xie, et al., ICASSP’19]

VLAD = Vector of Locally Aggregated Descriptors
Outline

Task Introduction

Speaker Embedding

End-to-end
Can we jointly learn speaker embedding and similarity computation?

Model

> threshold?  
Same

< threshold?  
Different

Positive Examples:

K Enrollment Utterances

K utterances from spk i

1 utterance from spk i
Can we jointly learn speaker embedding and similarity computation?

Model scalar > threshold? Same
< threshold? Different

Negative Examples:
K Enrollment Utterances
K utterances from spk i
1 utterances from spk j

Also refer to generalized end-to-end (GE2E) [Wan, et al., ICASSP’18]
End-to-end
[Heigold, et al., ICASSP’16]

<table>
<thead>
<tr>
<th>Table 1: Data set statistics.</th>
<th>#utterances (#augmented)</th>
<th>#speakers</th>
<th>#utts / spk</th>
</tr>
</thead>
<tbody>
<tr>
<td>train_2M</td>
<td>2M (9M)</td>
<td>4k</td>
<td>&gt;500</td>
</tr>
<tr>
<td>train_22M</td>
<td>22M (73M)</td>
<td>80k</td>
<td>&gt;150</td>
</tr>
<tr>
<td>enrollment evaluation</td>
<td>18k</td>
<td>3k</td>
<td>1-9</td>
</tr>
<tr>
<td></td>
<td>20k</td>
<td>3k</td>
<td>3-5</td>
</tr>
</tbody>
</table>
End-to-end
[Heigold, et al., ICASSP’16]

Text-dependent v.s. Text-independent

Enrollment Utterance 1

Enrollment Utterance K

Different Speakers

Evaluation Utterance

Network

Network

Network

average

Similarity

score
Text-independent

[Yun, et al., INTERSEECH’19]
Concluding Remarks

Task Introduction

Speaker Embedding

End-to-end
Reference

• [Variani, et al., ICASSP’14] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, Javier Gonzalez-Dominguez, Deep neural networks for small footprint text-dependent speaker verification, ICASSP, 2014


• [Snyder, et al., ICASSP’18] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, Sanjeev Khudanpur, X-Vectors: Robust DNN Embeddings for Speaker Recognition, ICASSP, 2018

• [Wan, et al., ICASSP’18] Li Wan, Quan Wang, Alan Papir, Ignacio Lopez Moreno, Generalized End-to-End Loss for Speaker Verification, ICASSP, 2018

• [Yun, et al., INTERSPEECH’19] Sungrack Yun, Janghoon Cho, Jungyun Eum, Wonil Chang, Kyuwoong Hwang, An End-to-End Text-independent Speaker Verification Framework with a Keyword Adversarial Network, INTERSPEECH, 2019
Reference


• [Chung, et al., INTERSPEECH’18] Joon Son Chung, Arsha Nagrani, Andrew Zisserman, VoxCeleb2: Deep Speaker Recognition, INTERSPEECH, 2018

• [Xie, et al., ICASSP’19] Weidi Xie, Arsha Nagrani, Joon Son Chung, Andrew Zisserman, Utterance-level Aggregation For Speaker Recognition In The Wild, ICASSP, 2019

• [Chowdhury, et al., ICASSP’18] F A Rezaur Rahman Chowdhury, Quan Wang, Ignacio Lopez Moreno, Li Wan, Attention-Based Models for Text-Dependent Speaker Verification, ICASSP, 2018