Speaker Diarization

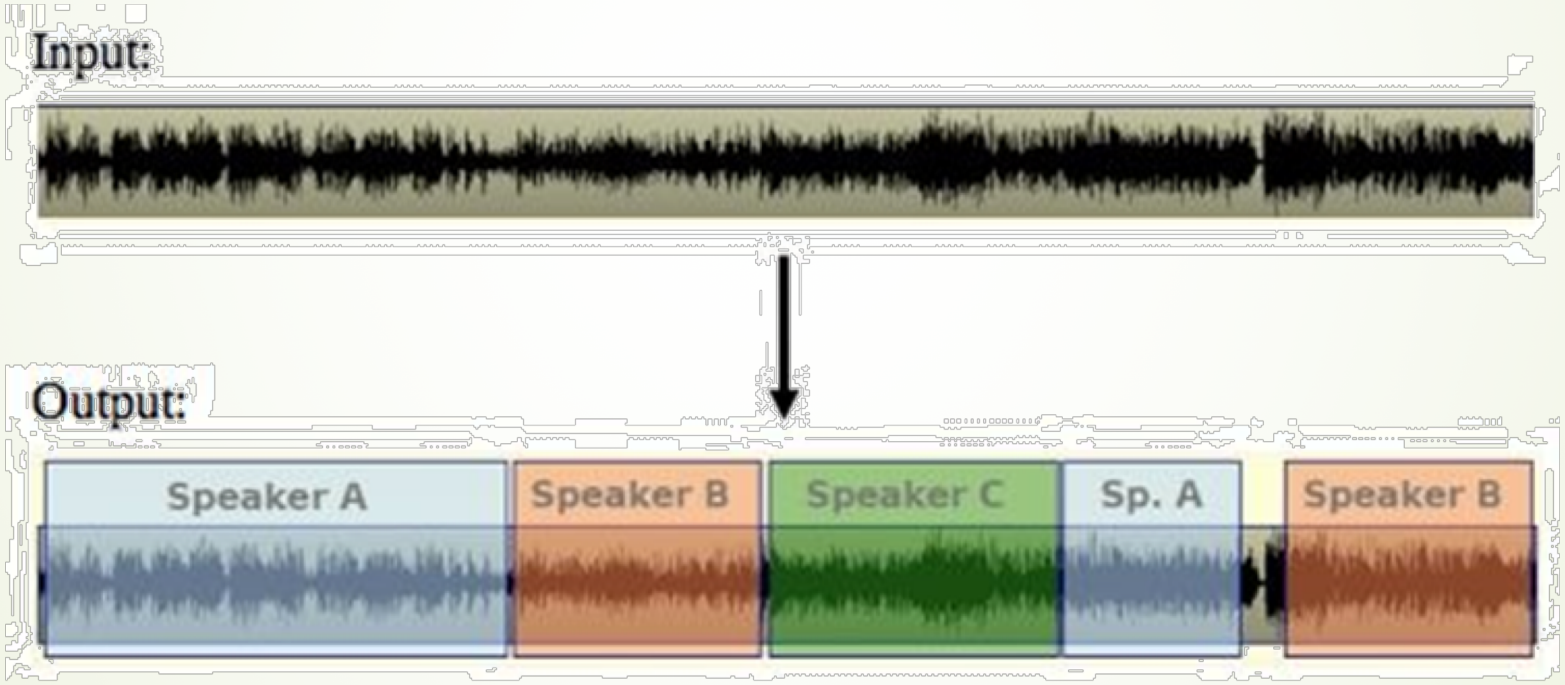
Speaker Recognition

- Speaker Recognition （说话人识别）可以分为说话人辨认（ Speaker Identification ）和说话人确认（ Speaker Verification ）。

- Speaker Verification （说话人确认）判断说话人的身份与其声明的身份是否是同一人。
- Speaker Identification（说话人辨认）根据说话人的语音确定其为待选的多个说话人中的某一个。

Speaker Diarization

- 研究内容 “who spoke when”



Speaker Diarization

基本步骤

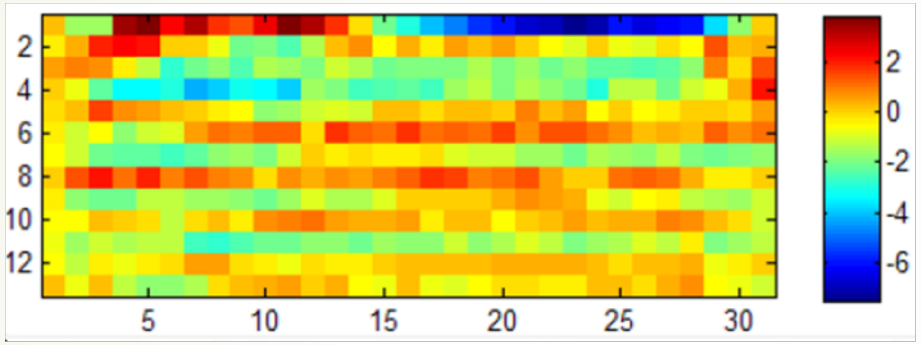
01 Audio embedding extraction：MFCCs, speaker factors, i-vectors, d-vectors

02 Speech segmentation：将输入的音频切分成只有一个讲话人的短段音频。

03 Clustering：对所有碎片段，进行聚类，把属于同一个说话人的片段都聚在一起。

现有提取特征算法

- Hand-crafted feature ：MFCCs

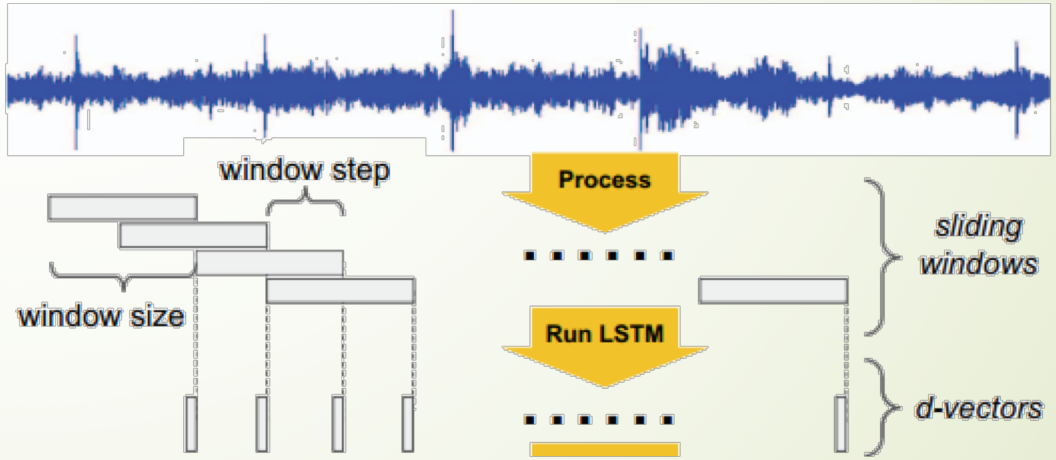


- 基于说话人确认(SV)模型：生成能代表说话人身份的矢量

GMM-UBM-JFA(Universal Background Model-Joint Factor Analysis,)

--- i-vectors

LSTM --- d-vectors



现有音频分割、聚类算法

- 分割算法

  基于静音检测 – 音量的大小

  基于距离度量

  基于模型搜索

- 聚类算法

  Gaussian mixture models, mean shift, agglomerative hierarchical clustering, k-means, and spectral clustering. Unsupervised

FULLY SUPERVISED SPEAKER DIARIZATION -- Google

- **Unbounded interleaved-state RNN**, a trainable model for the general problem of segmenting and clustering temporal data by learning from examples. Supervised framework

- Each speaker is modeled by an instance of RNN, and these instances share the same parameters;
- An unbounded number of RNN instances can be generated;
- The states of different RNN instances, corresponding to different speakers, are interleaved in the time domain.

https://github.com/google/uis-rnn

FULLY SUPERVISED SPEAKER DIARIZATION -- Google

- Generative process of UIS-RNN

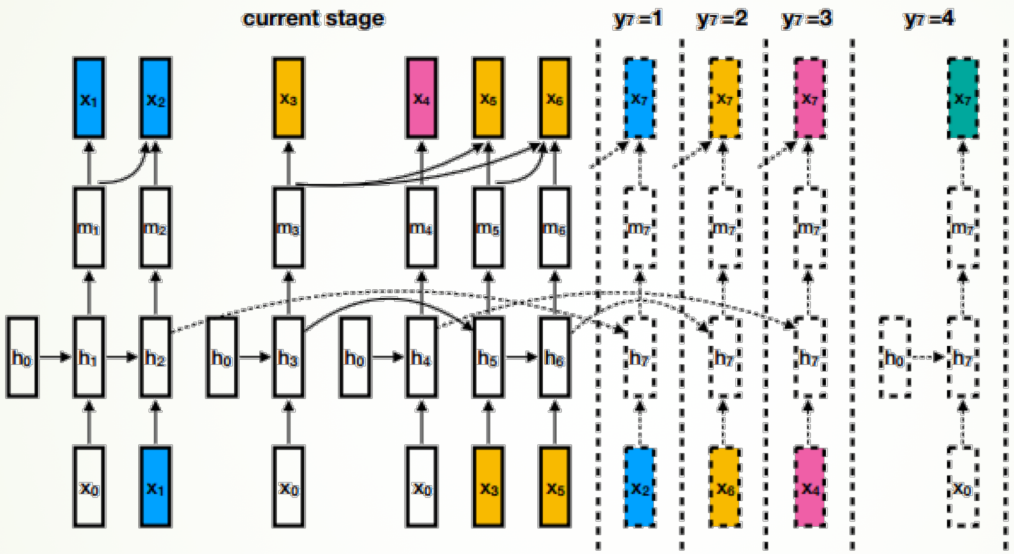


Fig. 2. Generative process of UIS-RNN. Colors indicate labels for speaker segments. There are four options for y_7 given $\mathbf{x}_{[6]}, y_{[6]}$.

```
Ground truth labels:
['15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1',
'15_1', '15_1', '15_1', '15_2', '15_2', '15_2', '15_2', '15_2', '15_2', '15_2', '15_2', '15_2', '15_2', '15_2', '15_2', '15_0', '15_0', '15_0', '15_0', '15_0', '
15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_2', '15_2', '15_2', '15_2', '15_2', '15_2', '15_2', '15_0', '15_0', '15_0', '15_0', '1
5_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15
_0', '15_0', '15_0', '15_0', '15_0', '15_0', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1', '15_1']
Predicted labels:
[0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2
, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

https://github.com/google/uis-rnn

FULLY SUPERVISED SPEAKER DIARIZATION -- Google

- Diarization Error Rate (DER) on NIST SRE 2000 CALLHOME

d-vector	Method	Training data	DER (%)
V1	k-means	—	17.4
	spectral	—	12.0
	UIS-RNN	5-fold	**11.7**
	UIS-RNN	5-fold + Disk-6 + ICSI	**10.6**
V2	k-means	—	19.1
	spectral	—	11.6
	UIS-RNN	5-fold	**10.9**
	UIS-RNN	5-fold + Disk-6 + ICSI	**9.6**
V3	k-means	—	12.3
	spectral	—	8.8
	UIS-RNN	5-fold	**8.5**
	UIS-RNN	5-fold + Disk-6 + ICSI	**7.6**
Castaldo *et al.* [4]			13.7
Shum *et al.* [9]			14.5
Senoussaoui *et al.* [10]			12.1
Sell *et al.* [1] (+VB)			13.7 (11.5)
Garcia-Romero *et al.* [2] (+VB)			12.8 (9.9)

谢谢