Pyannote评价体系

报告人：郑群

时间：2019年1月04日

内容简介：

- 简介：
- （1）语音活动检测（VAD）又称：语音边界检测，语音端点检测。通俗点说就是一段语音在哪些时间段内是有声音活动的，哪些时间段内是静音的。目标就是在音频流中找到音频变化点
- （2）Speech_Diarization:主要解决两个问题：[1]：who speak when : 什么人，在什么时间，说话了 [2]：说话者分析是说话者分割（segments）和说话者聚类(cluster)的结合。目标是基于说话者特征将语音片段组合在一起。
- （我目前在做的就是：who speak when）

Pyannote-metric

- 简介：
- Pyannote-metric：一种speaker diarization系统中的评价，诊断，错误分析工具。
- 操作流程：

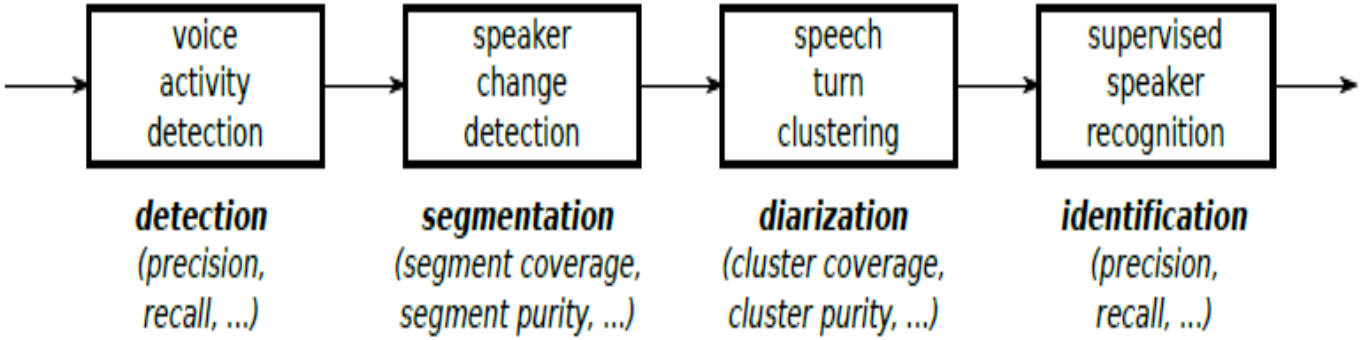

Figure 1: *A typical pipeline for speaker diarization, aligned with the list of available evaluation metrics*

Pyannote-metric

- 输出结果形式：

Diarization (collar = 0 ms)	error	purity	coverage	total	correct	%	fa.	%	miss.	%	conf.	%
BFMTV_BFMStory_2011-03-17_175900	14.64	94.74	90.00	2582.08	2300.22	89.08	96.16	3.72	80.14	3.10	201.72	7.81
LCP_CaVousRegarde_2011-02-17_204700	17.80	89.13	86.90	3280.72	2848.42	86.82	151.78	4.63	208.29	6.35	224.01	6.83
LCP_EntreLesLignes_2011-03-18_192900	23.46	79.52	79.03	1704.97	1337.80	78.46	32.89	1.93	157.14	9.22	210.03	12.32
LCP_EntreLesLignes_2011-03-25_192900	26.75	76.97	75.86	1704.13	1292.83	75.86	44.61	2.62	158.38	9.29	252.92	14.84
LCP_PileEtFace_2011-03-17_192900	10.73	93.33	92.30	1611.49	1487.32	92.30	48.73	3.02	55.49	3.44	68.67	4.26
LCP_TopQuestions_2011-03-23_213900	18.28	98.25	94.20	727.26	668.65	91.94	74.36	10.22	16.41	2.26	42.20	5.80
LCP_TopQuestions_2011-04-05_213900	27.97	97.95	79.81	818.03	638.68	78.08	49.45	6.04	17.46	2.13	161.89	19.79
TV8_LaPlaceDuVillage_2011-03-14_172834	21.43	92.89	89.64	996.12	892.04	89.55	109.36	10.98	11.80	1.18	92.28	9.26
TV8_LaPlaceDuVillage_2011-03-21_201334	66.23	77.24	70.64	1296.86	691.76	53.34	253.80	19.57	29.16	2.25	575.95	44.41
TOTAL	23.27	88.18	84.55	14721.65	12157.71	82.58	861.14	5.85	734.28	4.99	1829.67	12.43

- VAD计算公式：

$$\text{detection error rate} = \frac{\text{false alarm} + \text{missed detection}}{\text{total}}$$

- false_alarm:non_speech预测为speech_region

Pyannote_metric

- Miss_detection : speech_region预测为non_speech
- Total:真值中说话区域总时长
- Diarization计算方法：

$$\mathrm{DER} = \frac{\text{false alarm} + \text{missed detection} + \text{confusion}}{\text{total}}$$

- Confusion：说话人混淆区间：该部分计算是特别麻烦的。同时需要考虑标签和不同标签的分割段数长短

Pyannote-metric

- Confusion分析：作者在计算该段真值和预测值的重合区间时，需要考虑重合区间标签相同和不同时的情况。
- 比如：真值：[0,10] :’a’ 预测值：[0,1]:’a’
- [1,3]:’b’
- [4,5]:’c’
- [6,10]:’d’
- 按照作者的思想：d为[0,10]区间的另一说话人
- 其他三个为混淆

Pyannote-metric

- 真值：[0,10] :’b’ 预测值：[3,5]:’a’
- [6,7]:’b’
- [7,8]:’c’
- [8,10]:’b’ (√)

真值:[0,10]:’a’ 预测值：[2,4]:’a’

[6,8]:’b’

[8,10]:’c’

此种情况下哪个作为预测值的真值都可以，其余为混淆

Pyannote-metric

- IER: Identification Error Rate:
- 计算公式：

$$\text{IER} = \frac{\text{false alarm} + \text{missed detection} + \text{confusion}}{\text{total}}$$

- 这个公式跟DER是一模一样的额，含义不同
- Confusion：直接比对真值和预测值的label，而不是one-to-one matching。

总结：

- （1）Pyannote-metric：为计算VAD和Speech Diarization提供了接口，可以便于用户调用 。
- （2）Vad_error：计算上跟作者有点出入
- （3）Diarization：计算思路是正确的，可以拿来借鉴。至于性能是不是最好的有待商榷，可能别的论文中有更好的思路。

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

Thank You

请各位师兄师姐们指导批评