多模态说话人识别中的“星亮猜想”

音视频说话人识别基于语音和人脸两种信息来进行身份认证。一个问题显然是：这两路信号具有什么样的特性才有融合的意义？

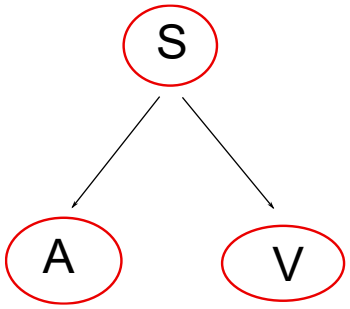
理论上，如果 A 和 V 组合起来与身份 S 具有更强的互信息量，则融合有意义。这等价于 H(X|A,V)比 H(X|A)或 H(X|V)更大。直观理解，意味着 A,V 融合的分类系统具有更好的分类效果。

星亮猜想

一个问题是，如果 A 和 V 互相独立，那么应该如何对 A 和 V 两个系统做融合？一个直觉是，如果 A,V 独立，则做分数融合即可。**星亮猜想说，即使在这种独立信息源的情况下，单纯的分数融合也不是最优的。**例如，如果我们知道一个人本身视觉是男性，但声音是女声，单纯用视觉和声音在各种的模态中区分度都不高，但组合起来的特殊性可以帮助我们更有效地判断这个人。

独立性假设

首先我们确定一个事实，如果 A-V 互相独立，则必然有 A 或 V 与 S 独立。这是因为如果 S 可以同时影响 A 和 V，则 A-V 必然通过 S 具有相关性。



这一结论意味着在星亮猜想里讨论的前提并不存在。如果 A-V 真的独立，则不必融合，因为必有一个模态与 S 无关。

事实上，我们能确认 A 和 V 同时与 S 相关，因为一个人的固有特性（如性别，年龄等）与声音和面容都有关系，即 P(A|S)<>P(A), P(V|S)<>P(V)。事实上，如果考虑到 A,V 共有的这些隐变量，可知 A,V 之间并不独立。然而，我们一般可以认为 A/V 在 **S 可见时**独立，即 P(A,V|S)=P(A|S)P(V|S)。我们考察这种条件独立下的打分情况。

1：N 辨认任务：

$$P(S|A,V)=\frac{P(A,V|S)P(S)}{P(A,V)}=\frac{P(A|S)P(V|S)P(S)}{P(A,V)}=\frac{P(S|A)P(S|V)P(A)P(V)}{P(A,V)P(S)} \quad [1]$$

在 1:N 任务中，如果 S 的先验相等，只考察分数的排序，因此只有 P(S|A)和 P(S|V)两项相关。故而只需对 P(S|A)和 P(S|B)进行有效建模，即可通过分数融合，得到最优解。

如果 S 本身是由另外一组 A/V 数据注册而成，则事情会复杂得多：

$$P(S|A,V)=\frac{P(A,V|S)P(S)}{P(A,V)}=\frac{\int P(A,V|X)P(X|A1,V1)dXP(S)}{P(A,V)}=\frac{\int P(A,V|X)P(A1,V1|X)P(X)dX\ P(S)}{P(A,V)P(A1,V1)} \quad [2]$$

$$P(S|A)P(S|V)=\frac{\int P(A|X)P(X|A1)dX\int P(V|X)P(X|V1)dX\,P(S)P(S)}{P(A)P(V)}=$$

$$\frac{\int P(A|X)P(A1|X)P(X)dX\int P(V|X)P(V1|X)P(X)dX\,P(S)P(S)}{P(A)P(V)P(A1)P(V1)} \quad [3]$$

同样假设 S 等先验，考虑和 S 相关的项，相关项包括：

$$P(S|A,V)\propto\frac{\int P(A,V|X)P(A1,V1|X)P(X)dX}{P(A1,V1)} \quad [4]$$

$$P(S|A)P(S|V)\propto\frac{\int P(A|X)P(A1|X)P(X)dX\int P(V|X)P(V1|X)P(X)dX}{P(A1)P(V1)} \quad [5]$$

可以看到，这时的统一建模与分数融合并不等价。

1：1 确认任务

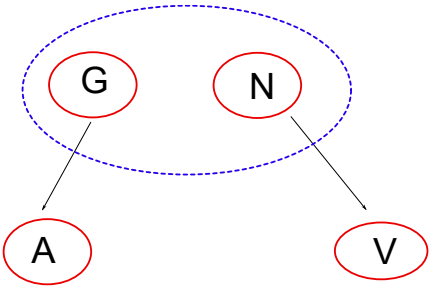
确认任务比辨认任务的打分需要加入一个正则项，对统一建模，这一正则项为 p(A,V)，对分数融合，这一正则项为 p(A)P(V)。因此，当注册数据较少时，参考 1：N 的结果，统一建模与分数融合显然是不相等的。

我们仅讨论当注册数据较多，P(X|A1,V1)可以用一个确定向量 S 来代表时，此时统一建模的 LR 为$\frac{P(A,V|S)}{P(A,V)}=\frac{P(A|S)P(V|S)}{P(A,V)}$，而分数融合的 LR=$\frac{P(A|S)P(V|S)}{P(A)P(V)}$。可以看到，在确认任务中，即便注

册数据足够多，分数融合依然不是理论最优解。

完全独立先验（星亮猜想）

实际情况下，如果 S 不止一个隐变量，而是一群，如 Gender(G)和 Age (N)，而 A 和 V 只与其中一个隐变量相关，如下图所示。



此时，A-V 是否独立，取决于群体中 N 和 G 的先验概率是否独立。如果先验概率独立，则 A-V 独立（如果不独立，则和前述讨论的情况相似），因此满足星亮猜想的条件。

我们以 1:N 任务来讨论。首先，因为 N,G 先验概率独立，当有足够多的 A/V 数据进行注册，则可以得到 A/V 的分别中心，uA 和 uV，其中 uA 只取决于 A，uV 只取决于 V。此时:

$$P(S|A,V)=\frac{P(A,V|S)P(S)}{P(A,V)}=\frac{P(A|uA)P(V|uV)P(S)}{P(A)P(V)}=\frac{P(S|A)P(S|V)}{P(S)}$$

此时分数融合适用。

如果 S 本身是由另外一组 A1/V1 数据注册而成，则:

$$P(S|A,V)=\frac{P(A,V|S)P(S)}{P(A,V)}=\frac{\int P(A,V|X)P(X|A1,V1)dX\,P(S)}{P(A,V)}$$

$$=\frac{\int P(A|uA)P(V|uV)P(A1|uA)P(S1|uV)P(uA)P(uV)duA\,duV\;\;P(S)}{P(A,V)P(A1,V1)}$$

$$=\frac{\int P(A|uA)p(A1|uA)p(uA)duA\int\;\;P(V|uV)P(V1|uV)P(uV)duV\;\;P(S)}{P(A,V)P(A1)P(V1)}$$

上式表明，当 A/V 真的统计独立时，分数融合是最优解。

同样的分析用于 1：1 确认任务中。

结论：

（1） A-V 可能存在条件独立，但本身并不独立。以往我们测试得到的较低相关性原因在于两者之间与 S 非相关的信息过大，导致无法检测出其相关性。
（2） 在 1：N 辨认任务中，如果注册数据较多，可用分数融合方法；但当仅有一条注册数据时，此时与分数融合方法并不等于理论最优。
（3） 在 1：1 确认任务中，分数融合并不是理论最优解。
（4） 假设 A-V 确实独立（如 V 中去掉性别信息，A 中去掉年龄信息等），则分数融合是最优解。星亮猜想不成立。