对英国关于法庭话者比较的立场声明的回应

Phil Rose *Geoffrey Stewart Morrison*

School of Language Studies, Australian National University

澳大利亚国立大学语言研究学院

philip.rose@anu.edu.au geoff.morrison@anu.edu.au

1 引言

最近一期的“the *International Journal of Speech Language and the Law（国际言语、语言和法律）*”期刊刊登了一篇题为“position statement concerning the use of impressionistic likelihood terms in forensic speaker comparison cases（关于利用似然率方法进行法庭话者比较案件的立场声明）”的文章（French and Harrison 2007）。这一立场声明是英国许多研究人员和法庭工作人员共同合作的结果。文章在前言中指出：

> “这一声明曾在英国所有的从事法庭言语研究的科学家和感兴趣的专业人士中传阅。除一人外，其他所有联系到的人都在此签了名。现在这一声明反映了英国几乎完全一致的立场。”(p. 138)

这一立场声明也被提交到了苏格兰、北爱尔兰、英格兰和威尔士等地的起诉机构。为了简便起见，我们将其称之为“*UK Position Statemen*（英国立场声明）”，附加说明的是这恐怕不能反映英国所有相关团体或所有司法机构的看法[1]。

*“国际言语、语言和法律”*期刊的编辑约稿在后续的期刊中发表对于英国立场声明的回应。该文就是对这一立场的回应，本文的初稿曾在2008年7月的第17届法庭国际语音学和声学学会上宣读。

我们首先按照自己的理解总结了英国的立场声明，然后陈述我们的观点。这样有助于阐述我们认为是正确的提供法庭语音比较证据的方法体系。

2 英国立场声明的说明

英国立场声明的前言中已经清楚地表明，所谓的语音比较指的是由专家进行的比较，因此是技术比较，而不是无经验的比较(关于这些术语参见 Nolan 1983: 7, 1997: 744–745)。

2.1 动机和目标

[1]因为我们知道，在英国有两个有兴趣的专业人士的意见没有得到征询，而不是像文章中说的那样，它代表了“所有法庭言语科学家和所有感兴趣的专业人士的意见。”。

在前言中，英国立场声明指出，他们的动机是出于对法庭话者比较案件中结论的表述方法的考虑（p. 13）。据说，意识到现存的方法体系有问题，最早出现于英格兰和威尔士的上诉法庭，在裁定*R v. Doheny 和 Adams* ([1996] EWCA Crim 728)案件中，牵涉到与一名DNA专家提供的主要证据相关的错误起诉推论。

英国立场声明的前言称：

> “…有一种新方法带来了分析家和证据角色的根本改变。在过去，法庭言语学家的任务经常被认为是鉴定说话人。在新方法中，他们的角色不是做鉴定，而是要评价涉案录音与嫌疑人语音是否相符。(p. 138)”

英国立场声明的注脚2补充说，他们进行的不是鉴定，而是比较。前言中也说到，提出英国立场声明的目的是“…为了取得与其它法庭科学领域现代思想的一致。(p. 137)”，并且，“这一新的方法体系，从概念上讲，与现在的DNA证据是一致的。(p. 138)”

文章的结尾，英国立场声明的作者和签名者说：

> “…原则上接受并愿意考虑在似然率（包括贝叶斯理论）概念范围内进行法庭话者比较。然而，考虑到缺少定义相关背景人群的人口数据，目前还不便于应用这种定量化方法。”(p. 142, §6)

可以看出，英国立场声明是在合理地回避基于似然率的方法体系，而退而求其次地支持另一种主张。然而，这是不公正的。我们可以清楚地看出，尽管没有明说，实际上他们试图将基于似然率的方法折中，而这其实也是提出这项主张的深层动机。我们要说的是，这样的折中并不是最好的选择，从概念上讲，这与DNA证据体系也不对等。

2.2 英国理论体系

图1表示的是英国立场声明中提出的理论体系的流程图。在英国理论体系中，语音样本的比较依次分为两部分：一致性比较和特殊性比较。

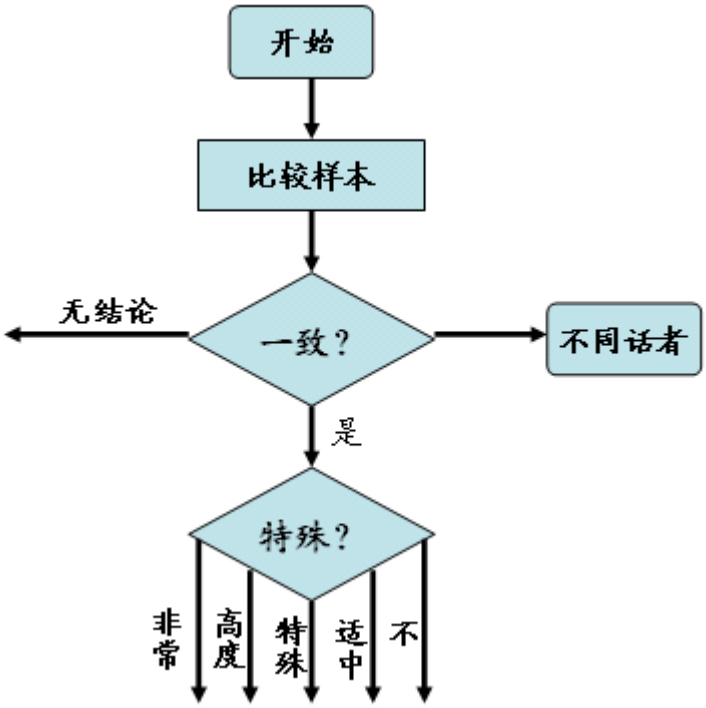

图1　英国分析体系的流程图

2.2.1　一致性

一致性的特点是“样本语音和检材语音是否符合或一致，是否来自同一个发音人。”(p. 141, §4.1)。它的评估是利用“能够观察到的特征的相似或差异程度” (p. 141, §4.1)。样本和检材之间的差异如果从声学模式、语音或语言的变异（例如：不同的信道特征、或来源于话者自身的社会语言学、心理学和/或生理因素）等方面解释不通的话，就可以算做不一致” (p. 141, §4.1)。一致性的量化程度分为三个级别：一致，不一致和无结论。如果比较的结果是不一致，那就意味着两份样本来源于不同话者。如果比较的结果是一致，下一步就要考虑特殊性问题（这样，一致性和特殊性是有先后次序的，特殊性评断只有在确定为一致的前提下才能进行。）至于什么情况下属于无结论，文章没有说明。

2.2.2　*特殊性*

英国的立场声明强调确定为一致并不是说已知的和嫌疑样本肯定是一个人说的，因为“导致一致决定的一组特征…可能是许多人共有的特征” (p. 141, §4.2) 。这就意味着如果这组特征是与众不同的或者不同寻常的，那么它们来自于同一话者的样本的似然率会更大。特殊性的评价从无特色到特别有特色分为5个层次，对后者的解释是：“其他人具有同组特征的可能性是极其微小的。” (p. 141, §4.2)。文中也没有说明如何确定样本的特殊性。我们假定这是准备由专家报告两个样本的特征一致，指向同一个人的发音，并且提供样本的特殊程度表明假如两者的一致性不是来自于同一人，那么它们的独特程度如何。

3 对英国立场声明的回应

首先，我们为英国立场声明的动机鼓掌，而且乐于接受其总的意向。特别要指出的是，我们完全赞同将法庭语音比较证据与法庭科学其它领域的现代思想取得一致这一目标，特别是将其与DNA证据的评价应用取得一致。这些目标并不是新的话题，一些学者也提出过同样的主张。（Champod and Meuwly (2000), González-Rodríguez et al. (2006), González-Rodríguez et al. (2007), Rose (2002, 2003), Saks and Koehler (2005)）但是，我们要说的是英国立场声明并没能达到这样的目标。

在批判地分析英国立场声明以前，我们首先提出我们认为不论从逻辑上还是法律上都是正确的法庭比较证据的评价体系：即似然率体系。我们相信这一体系代表了法庭科学其它领域的现代思想，它的应用典范就是当前的DNA证据评价。法庭统计学家、法律专家和法庭科学家们撰写的许多教材、文章和评论都对这一立场提供了支持，如Aitken and Taroni (2004), Balding (2005), Champod and Meuwly (2000), Evett (1991, 1998), Friedman (1996), Good (1991), González-Rodríguez et al. (2006), Haigh (2005), Hodgson (2002), Lindley (1991), Robertson and Vignaux (1995)等。

3.1 似然率体系（理论框架）

以下论述摘自Rose 的文章(2005: 49–54)。尽管该文章是以语音样本比较为对象的，但是原则上它适用于任何法庭证据（如：DNA、指纹、弹道和工具痕迹等等）。它们都需要将已知样本（样本）和涉案样本（检材）进行比较，并且样本的物理特性可以量化，样本之间的特性也可能会产生变化。如果您想更全面了解似然率理论体系，可以参看以下论著：Balding (2005), Lucy (2005), Robertson and Vignaux (1995), Rose (2002, 2003)。

在进行法庭语音比较时，典型的情况就是将一份未知的语音录音（通常来自于罪犯）与一份或多份已知的语音录音（通常来自于嫌疑人）进行比较。感兴趣的相关团体（如：警察/起诉，调查官）想知道是否未知语音和已知语音来自同一个人。通常他们明白：这得不到肯定的答案。毕竟，审判是对不确定事件做判断。所以，他们通常会问：两套样本是同一个人所说的可能性有多大？—这是很合乎常理的问法，因为哲学家和统计学家都认为衡量不确定事件的最好的方法就是使用概率(Lindley 1991)。当然，这其实也是暗指证据的作用。确切地说，实际问题就是：在基于证据的前提下，涉案样本和已知样本是同一人所说的概率有多大？传统而简便的条件概率表达公式如（1）所示，

$$p(\mathrm{H}_{ss} \mid \mathrm{E}_{sp}) \tag{1}$$

这里，“ p ”代表概率，“H_{ss}”代表同一话者的起诉假设，竖线“|”代表假定或者基于此条件，“E_{sp}”代表语音证据，即嫌疑人与被告语音样本之间的不可避免的差异。通常采用分数形式代替概率形式，因此，正式的表达式如（2）所示，

$$p(H_{ss} \mid E_{sp}) / p(H_{ds} \mid E_{sp}) \tag{2}$$

这里，“H_{ds}”代表样本来自于不同话者的假设。

这一公式来源于十八世纪中期的贝叶斯定理(Bayes, 1763)。贝叶斯定理的重要之处就在于它可以提供证据支持假设的概率，由此决定了它在法庭鉴定方面的特殊地位。贝叶斯定理非正式地表明可以从两个方面估计出证据支持假设的概率：一个是引入证据之前假设的概率，另一个是证据的力度。

适用于法庭语音比较的贝叶斯定理的分数形式如（3）所示，等式左边是语音证据支持同一话者假设的比值，即众所周知的后验比（*Posterior Odds*），它等于支持同一话者假设的先验比（*Prior Odds*）乘以证据的力度（*Likelihood Ratio*）。这样，支持同一话者假设的比值可以通过两个分项算出来，即先验比和似然率。

$$\underbrace{\frac{p(H_{ss} \mid E_{sp})}{p(H_{ds} \mid E_{sp})}}_{\textit{Posterior Odds}} = \underbrace{\frac{p(H_{ss})}{p(H_{ds})}}_{\textit{Prior Odds}} \times \underbrace{\frac{p(E_{sp} \mid H_{ss})}{p(E_{sp} \mid H_{ds})}}_{\textit{Likelihood Ratio}} \tag{3}$$

先验比是语音证据引入之前支持起诉假设的比值，仅仅是同一话者的概率除以不同话者的概率。就其极限讲，留下语音的可以是世界上的任何人，但是考虑到性别、口音和其它实际情况等明显的信息，先验比通常可以缩小到一定范围内。

似然率是法庭语音比较中最重要的度量标准，因为它能衡量支持假设的证据力度，这也正是专家应该评估的内容。公式（3）表明似然率本身也是概率的比值，但是这些概率是证据的概率，不是假设的概率。似然率量化的是：基于嫌疑人和罪犯的语音样本的差异，假定它们来源于同一话者的概率比来源于不同话者的概率大多少。

如果根据语音证据，两套样本来源于同一话者的可能性大于来源于不同话者的可能性，即如果$p(E_{sp} \mid H_{ss})$ 大于 $p(E_{sp} \mid H_{ds})$，这就支持样本来源于同一话者的起诉假设；相反，如果根据语音证据，两套样本来源于不同话者的可能性大于来源于同一话者的可能性，即如果$p(E_{sp} \mid H_{ds})$ 大于$p(E_{sp} \mid H_{ss})$，那就支持辩护假设；如果两套样本来源于同一话者的可能性和来源于不同话者的可能性相当，即如果$p(E_{sp} \mid H_{ss})$ 和 $p(E_{sp} \mid H_{ds})$的比值是1，则证据无效。

这样，似然率的大小就量化了证据的力度：其值大于1表明支持同一话者主张，小于1表明支持不同话者主张，等于1（或接近1）表明证据无效（或几乎无效）。

有关法庭证据评价(如：Robertson and Vignaux, 1995)和法庭统计(如：Aitken and Stoney 1991, Aitken and Taroni 2004, Lucy 2005)方面的主要教材强调了法庭专家的任务是通过估算似然率来量化证据的力度，即在起诉和辩护这两个竞争假设下证据的概率。他们写道：

“采用该方法分析案件，不管对象是DNA、玻璃碎片、衣物纤维还是其它任何客体，都是完全适用的。” (Haigh 2005)

“统计评价，特别是贝叶斯方法，诸如似然率的计算…相比其它凭直觉或者猜测来说，是目前唯一的、正确合理地量化证据的方法。” (Lucy 2005: 138)

3.2 是比较，不是鉴定

英国立场声明强调：他们主张对语音样本进行比较，而不是鉴定，这是他们采用的术语（我们已经接受）。这是值得一提的，因为现在有很多术语都在使用：如：辨认(*identification)*, 识别(*recognition*)、确认(*verification*)、鉴别(*discrimination*)等。部分文献中将辨认和识别混淆使用；而在其它部分文献中，辨认和确认被用来指逻辑上不同的两种分析方法和应用方法。我们同意英国立场声明中“比较是最合适的术语”的观点，但是原因略有不同。诸如辨认（*identification*）、确认（*verification*）和识别（*recognition*）等术语实际上指的是后验概率（即嫌疑人和罪犯语音相同的概率），有一些还有确定类别的含义。正如我们下面讨论的那样，Rose也曾论述过（2002: 89），让法庭专家提供后验概率，逻辑上是不可能的，法律上也是不合适的，因此与辨认（*identification*）、确认（*verification*）和识别（*recognition*）无关，我们同意英国立场声明中关于法庭工作中应该避免使用这类术语的观点，支持采用“比较”这样中立的术语，没有确定后验概率的含义。我们还建议，既然比较的对象是语音录音，也就是说比较的是语音，而不是说话人，因此我们采用“法庭语音比较（*forensic voice comparison*）”这一术语，而不是英国立场声明中的法庭话者比较（*forensic speaker comparison*）。Nolan(1983, 1996)针对语音在法庭的应用曾给予迄今为止最好的描述。Rose (2002: Ch. 10) 在他的文章中也曾详细说明了他的符号语言。

3.3 禁止使用基于证据条件下的假设概率

英国立场声明中最重要的提议就是专家应避免给出基于证据条件下的假设概率[p(H|E)]，我们举双手赞成。这也是当今法庭统计学家们一直采取的立场和近来越来越多的法庭采取的立场（参见Aitken and Taroni 2004: 108, 122–128, 153–155, 208–213, Balding 2005: 145–153）。然而，英国立场声明暗指法庭科学家之所以不该引用p(H|E)的原因是由于它对证据的虚假评价造成起诉谬误。起诉谬误是指证据和假设的错位，即错误地用p(H|E)代替了p(E|H)。这就等

于说：由于证据支持有罪假定的概率是支持无罪假定的1000倍，因此被告有罪的概率是无罪概率的1000倍(Aitken and Taroni 2004: 79–82, Balding 2005: 146–147, Donnelly 2005, Evett 1998, Thompson and Schumann 1987)。毋庸置疑，法庭科学家本身应该避免这样的错误，应该尽一切努力防止律师、法官和陪审团等团体对专家证词的错误解释。但是，英国立场声明中关于“虚假评价”的争论并没有强调为什么法庭科学家必须提供基于假设条件下的证据概率[p(E|H)]，而不能提供基于证据条件下的假设概率[p(H|E)]。

为什么法庭专家不能提供基于证据条件下的假设概率的原因有两个：一是逻辑上的原因，二是法律上的原因。逻辑上的原因一般说来，就是遵循贝叶斯定理：后验比决定于证据力度（似然率）和先验比（参见公式3）。专家不知道先验比，因此逻辑上无法引用后验比。法律上的原因与最终裁定原则有关：在罪犯的样本确实表明有罪的案件中，专家宣称罪犯的言语可能是嫌疑人所说，这无异于表明嫌疑人可能有罪，这样就等于侵占了事实裁定者（是法官还是陪审团取决于法律体系）的角色。

虽然英国立场声明似乎是在谴责提供基于证据条件下的假设概率，但是实际上文章中有两处是建议提供p(H|E)的。这一点将在下面分两个部分进行讨论：一是关于DNA与语音样本的差别，另一个是关于闭集比较。

3.3.1 DNA与语音样本的差别

违背禁止引用p(H|E)原则的第一处出现在关于一致性问题上的“不一致”论述：

“当样本不一致时，作出样本为不同话者所说的论断在逻辑上是没有缺陷的。根据数据的情况可以给出自信程度。”(p. 141, §4.3)

如果说，因为样本不一致而给出一个样本为不同话者所说的自信度，这实际上就是p(H|E)陈述。如果按照英国立场声明中所说的去做，那么，根据贝叶斯定理，这事实上就是逻辑缺陷。

我们怀疑这种不一致已经慢慢渗透进来了，因为英国立场声明的作者试图去适应DNA的分析模式，但是他们没有考虑到DNA与语音证据在特性上的重大差异。虽然法庭语音证据的评价确实可以采用与DNA同样的方法进行，即采用似然率方法来进行—最近的一篇文章(González-Rodríguez et al. 2007)通过自动和传统的方法已经证明了这一点。但是，由于变异特性的存在，将法庭语音数据比照DNA时必须谨慎。对法庭而言，有三方面的变异是最重要的：即变异的类型、需要考虑的变异的水平以及变异的程度。DNA在这三个方面都与语音不同，但最重要的不同是变异的类型和水平。

变量可以是连续的或离散的，也可以是二者的组合。DNA变量，如：典型的特定位点上STR等位基因的长度是连续的。对离散变量，可以谈及相配，例如两个样本在基因型上显示相同的组合：D18上的14、16和9.3，9.3在THO1上(Balding 2005: 3) [9.3, 虽然看上去象连续的，实则不然。它表明一次重复只出现于三个碱基而不是四个碱基。]DNA匹配不上也是可能的。

DNA证据是离散的，而语音证据是连续的：倒谱系数，共振峰中心频率等等都是连续数值的变量，甚至更高水平的特征，如：特定音位的出现率也会出现连续数值的比例。此外，语音的特性随不同场合变化，而有机生物的DNA的测量则每次都是一致的（排除测量误差、污染、体细胞的变化、移植和嵌合体等因素。）。

那么，需要补充说明的是，分类和不变属性意味着：如果两个DNA数据不匹配，那么，假定它们来源于同一有机体的概率就是零。在这种情况下，似然率的分子是零，不管先验概率是多少，它们来源于同一有机体的后验概率都是零。因此，DNA可以用于提供明确的排除证据。但是，语音不行。一般说来，语音数据就其本质而言，不能给出这样明确的排除结论。可以想象，在特定的条件下，语音比较也可以得出明确的排除结论，如一个幼小的孩童的声道不可能产生典型的成人男性的较低的共振峰，但是这种情况下，两个语音听起来会明显不同，向法庭专家咨询这样的案件是几乎是不可能的。

再次补充说明的是，假定两个DNA数据匹配，那么匹配的两个样本来源于同一有机体的概率是1(Aitken and Taroni 2004: 404, Evett 1998)。似然率的分子是1，它的大小就取决于分母的大小。分母是随机匹配概率（在英国立场声明中指随机发生率[2]），即从相关人群中随机抽取的样本与已有的DNA样本匹配的概率。在这种情况下，因为似然率等于随机匹配概率的倒数，所以DNA证据的力度就可以直接用随机匹配概率来表示，而不需使用似然率。

可能是由于英国立场声明中不恰当地引用了DNA分析，因而使其陷入随机匹配概念问题。英国立场声明考虑到这样一个事实：在英国，嫌疑人和罪犯的DNA匹配已经明确，随机匹配概率为一百万分之一（也就是说，一百万个人中有一个人与罪犯的DNA匹配）。英国有六千万人口，由此，英国立场声明引用说："…有六十分之一的几率该DNA来源于被告。" (p. 139). [3] 英国立场声明继续说到：

---

[2] "由法庭引入的'随即出现率'这一术语[*R v. Dohney and Adams*]看起来与匹配概率是同义的。然而，这一新奇的创造并不受欢迎，因为现存的术语已有很多。对其不熟悉将会导致滥用。" Balding (2005: 152)

[3] 正确的答案实际上是被告是DNA痕迹来源的概率为1/61（用分数形式则为1/60）。在英国，可能是罪犯的6千万人中，只有一个人是真正有罪的，其余的59，999，999人都是无辜的。有罪一方将提供一个匹配结果，其余的59，999，999人将提供60个匹配结果。（因为随即匹配概率为 1/1,000,000, 所以59,999,999 × (1/1,000,000) = 60 (取最近的正整数)。所以，总共会可能有61个匹配。在这61个匹配中，只有一个是真正的匹配，其它的都是假阳性，所以嫌疑人留下痕迹的概率为1/61。Balding (2005: 11)给出了在这种情况下计算有罪概率的简化公式：[$P(G|E) = 1/1+N*p$]，其中，$P(G|E)$代表基于证据条件下有罪的概率，N是不包括嫌疑人在内的所有可能是罪犯的人数，p是随即匹配概率。

> “估计一百万人中有一个人会有相同的DNA被认为是它的‘随机发生率’。语音学家可以算出随机发生率的语音特征极少。只有基频（音高的度量值），发音速率（讲话的速度）和口吃等。”(p. 140, §3)

语音数据本身是连续变量，说话人的两次发音永远不会完全相同也已是老生常谈，语音样本之间的变异总是存在的，因此法庭语音比较的似然率的分子永远也不会是0或者1。随机匹配这一概念不适用于连续变量的语音数据。因此，基于连续数据的法庭语音比较的证据力度只能用似然率来表示。

随机匹配概率对于本身是连续值的特征，如基频，肯定是毫无意义的概念，而值得探讨的是，对于口吃等特征的出现率是可以算出随机匹配概率的。但是，这只有在假定具有习惯性口吃的人说话时总是出现口吃，而一般不口吃的人总也不出现口吃的条件下才可以。

3.3.2 闭集比较

第二个违背禁止使用$p(H|E)$的地方出现在英国立场声明中的第五部分：

> “然而，在少数案件中，独立证据显示已知说话人闭集是存在的，而且他参与了谈话。在这样的案件中，比较任务就成为“谁说了什么”的问题。在这种情况下，如果其语音足以和另一个人区分开，那么，我们认为作出明确的鉴定结论是对的。”(p. 142, §5)

由于差异足够大而作出明确的鉴定结论属于根据证据给出的假设概率的论断，逻辑上违背了贝叶斯定理。从证据效力的角度出发，闭集比较同样可以作为开集比较来对待（参见Rose 2002: 64, 74）。

3.4 两阶段评价

英国立场声明主张的另一个重要部分，而且是沿着正确方向迈出的受欢迎的一步，是它的一致性和独特性的二分评价。英国立场声明在4.2部分提到：证据价值不仅取决于两个样本的相似性，还取决于它们的典型程度。这一点是对的。两个既很相似又很典型的样本支持同一的证据价值就不如两个很相似但非典型的样本价值高。这一点并不总能被人理解，仅仅依据相似性就认定同一的情况并不少见。因此，英国立场声明能够澄清这点是可取的。

咋看英国体系中的一致性和特殊性术语好像和3.1部分讨论的似然率的分子和分母是对应的。然而，英国体系中对于一致性和特殊性的评价与似然率的计算并不对等。似然率必不可少的特征是分子和分母要采用相同的标准进行测量（它们都是概率密度值），而且相互直接相关（即以比率形式）。在英国体系中，一致性和特殊性是连续有序的，测量的标准也不同（一个是三个离散水平，一个是五个离散水平），二者也不是直接相关。

事实裁定者需要知道语音样本之间的差异是否更可能来源于同一话者或不同话者，或者不管是来源于同一话者还是不同话者，其可能程度是否相等。除非两个术语都以相同的标准量化并且直接相关，否则这样做是不可能的。因此，我们认为这是英国体系的两阶段评价的一个不足。

事实上，英国体系的一致性和特殊性的两阶段分析使人想起了Evett(1977)用于证据评价的两阶段术语：比较（*comparison*）和显著（*significance*）。在这种方法中，按照事先的标准首先决定是否相配，即两个样本都在三个标准差以内。然后，根据已发现的在相关人群中的相似程度进行评价(对这一体系的批评参见Aitken and Taroni 2004: 10–11, Evett 1991)。虽然两阶段评价的几种变体历史上曾经应用于DNA证据，但是现在已被似然率所取代，两阶段方法并非是现代应用实践的代表(关于DNA证据在这方面应用的解释的历史评论参见Foreman et al. 2003)。

英国体系的两阶段评价除了存在系统问题以外，这两部分的结构也有问题。下面我们讨论边缘效应问题和多变量数据问题。这两个问题都与英国体系中一致性和特殊性有几个有限的归类结果这一事实有关。我们讨论的是与一致性相关的问题，但是应该清楚的是，这也适用于对特殊性的讨论。三个归类也好，五个归类也罢，对于我们争辩的逻辑来说无关紧要，我们仍选择“一致的”这一不幸的词语进行讨论。

3.4.1 边缘效应

语音样本之间总是存在差异的，即使是同一个人在同一场合间隔几秒钟重复同一发音，这是语音学事实。此外，这些差异是渐变的，不是分级的，样本差异本身也不适于用英国体系的三种归类结果来表示。

元音共振峰可以看作是典型的渐变特征。这样的情形很常见，即比较来自于嫌疑人和罪犯的一套元音的共振峰的中心频率分布。表1给出的是从两段窃听电话谈话中提取的单个元音的共振峰频率的均值和标准差。为了便于论证，我们假定特征的分布足以接近正态分布，以保证均值和标准差适于代表集中趋势和离散度。

表 1. 两段不同的窃听电话谈话中澳大利亚英语 /əː/的15个发音的共振峰中心频率的均值和标准差（单位：Hz）

	F2	F3
嫌疑人		
均值	1429	2298
标准差	30	67
罪犯		
均值	1450	2329
标准差	48	56
均值差	21	31

把F2看作是一个单个特征。我们可以想象：即使没有具体的人口数据，多数法庭科学工作者也会根据他们的经验同意，在同一发音人不同场合正常发音时，央元音的F2均值变化21Hz（相差1.5%）是在正常的变化幅度内的，这是可以预见到的。如果按照英国体系，应该将它们归为一致判决。由于英国体系并没有详细说明如何才能确定为一致，我们只能假定是依据明确的变异量化结果来进行判决，如特征的标准差。表1中的数据表明，嫌疑人和罪犯的语音样本的F2的标准差都大于它们样本的F2的均值差。那么，从哪一点上可以确定特征的观察值是一致的，进而确定样本是一致的，是来源于同一话者呢？边界点应该是两个标准差，还是三个标准差？是否应该采用频率统计测试，如T检验，检验水平规定为0.05，还是0.01？这样的方法属于强加归类，伴有边缘效应(Robertson and Vignaux 1995: 118)。如果采用嫌疑人数据的标准差为标准，将边界点设定为两个标准差，F2均值相差59.9 Hz将被判定为一致，而相差60.1 Hz则将被判定为不一致（或者二者都判为无结论）。难道决定仅仅取决于这0.2Hz的差异吗？我们主张用于法庭语音比较的度量相似性的任何标准都应该是渐变的，而不是分类的。

3.4.2 多变量数据问题

关于英国体系的一致性的进一步的问题是比较样本时必须是多个特征的比较。英国立场声明在前言中将法庭语音比较的过程描述为：

> “[包括]将样本拆分为语音学和声学的组成要素（例如：音质、语调、韵律、节奏、调音速率、辅音和元音发音），然后分别进行分析。”(p. 138)

这里没有问题，因为多维性是有利于提高法庭语音区分能力的一个方面。然而，声明中并没有说明最终是如何将个体证据从每个要素中提取出来再合并到一起的。再回到表1中的共振峰数据和以嫌疑人数据的标准差为基准确定的两个标准差边界的例子（不管使用的是什么特征，也不管规定的阈值是多少，该论据同样适用。）如果F2均值相差50Hz，在边界线内；F3的均值相差140Hz，在边界线以外。那么，这个“一致性”如何决定？如果一对样本在九个特征上判定为一致，而在一个特征上判定不一致，怎么办？再有，英国体系有义务详细说明在这种情况下应该如何处理。在这种情况下，举例说明该方法应该是有效的，因为现在看来这一方法很难执行。如果要避开分类量化，只要采用多变量渐变方法来度量相似性即可，如多变量似然率的分子。文献中已经阐述了在似然率框架下处理多变量数据的许多方法(如：Aitken and Lucy 2004, Pigeon, Druyts, and Verlinde 2000, Reynolds, Quatieri, and Dunn 2000)。

3.4.3 “一致的”的语义学含义

英国体系的4.1 部分将一致性定义为：“观察变量相似或不同的程度” (p. 141, §4.1)。毫无疑问，这与估计涉案语音与已知语音在特定特征上的相似程度是一致的。然而，按照现在的情况，人们对一致性的认识很有限，它的执行是有问题的，还远远没有弄清楚。特别是，选择“一致性”这一词语来代表这一参数并不恰当。Robertson 和 Vignaux 在他们关于证据评价的论著中强烈批评了法庭专家的这种用法。

> “其中，最差的就是‘一致的’这一词语，它不幸地被法庭科学家、病理学家和律师们普遍使用……不幸的是它不能清楚表达……律师通常把‘与…一致’解释为‘适度强地支持’。” Robertson and Vignaux (1995: 56)

我们怀疑，陪审团，甚至是法官，都有可能将“与同一话者的特征一致”理解为“可能来自同一话者”。虽然我们清楚，在英国体系内，这并不是“一致性”的本意。Robertson 和 Vignaux也指出，就其正确意义讲，“一致的”几乎没什么认识论上的效力，因为“与假设一致”并不是指假设为真的可能性（事实与假设一致，但是不可能仅来源于“一致性”定义的“假设与特征一致”）。

3.5 人口和样本问题

如上所述，英国立场声明在3.1部分提到：对这类证据进行量化的似然率方法评价，不管多么适合，都是不可能实现的，主要有两方面的原因：即“如何定义参考人群的问题”和“缺少人口数据的问题”(p.142 §6)。我们愿意承认这些确实是问题，恐怕也是当前最紧迫的问题。第一个属于理论性问题，与采样的相关人群的选择、样本的大小有关；第二个是实践性问题，与实际收集的数据有关。下面，我们要依次说明这两个问题。我们要说，这些问题虽然现实存在，但是并不妨碍似然率方法的应用。

3.5.1 合适的参考人群

英国的立场声明提到：基于似然率的分析方法存在一个问题，即“定义相关参考人群”问题。采用似然率方法估算证据的效力需要有相关人群的参考人群样本或背景人群样本[4]。英国立场声明意识到这是一个问题，这无疑是正确的。这不仅是语音面临的问题，对于DNA样本的评价来说也仍然是个挑战。严格地讲，选择合适的人群进行取样决定于选择假设。在典型的案件中，起诉方通常会主张涉案语音与已知的样本语音相同，也就是说，两段语音均来自被告。辩护方会主张涉案语音不是被告的语音。然而，不可能简单地主张“涉案语音的说

[4]不幸的是，在文献中，“参考或背景人群”这些术语有时被用作“样本人群”。实际上，参考或背景人群与可能是罪犯的所有人群并不是一回事儿（后者可以用于估算先验比）。

话人不是被告，是其他人。”而更可能的主张是“涉案语音是与被告讲同一语言、同一方言、而且是同一性别的另一人所说。”辩护假设也可以更为严谨：涉案语音是（如警察和律师等一般人员）听起来与被告声音很像的某个人所说。或者，也可以说是，涉案语音来源于被告的兄弟或双胞胎兄弟（见Lucy 2005: 129–133，关于人群限制的讨论。）。在进行分析的时候，法庭语音比较专家并不知道具体的辩护假设，因此必须预测可能的辩护假设，因此，相关的参考人群也必须因个案而定(见Evett 和 Weir 1998: 44–45; Aitken 1991)。

如何进行参考人群的取样是常见问题，也是英国立场声明中含蓄指出的，即样本究竟应该取多少为宜？不管是从研究的角度，还是从法庭应用的角度，这都是非常重要的问题。答案就是：它取决于要求的精确度。Aitken在1991 年曾谈到解决这一问题的一些方法。

3.5.2 缺少分布信息

根据以下引证，我们认为：缺少人口数据，是指缺少典型的法庭比较特征在人群中的分布信息：“对于案件检验中的所有主要的语音和言语特征，我们只是不知道它们在人口中的分布情况。” (p. 140, §3)

英国立场声明在回应Rose(2002: 78)时指出了由此产生的后果：如果不知道特征在人群中的发生率，就不可能引用其在人群中分布的随机概率。因此，“法庭语音学家就无法提供概率的统计数据。” (p. 140, §3)

很明显，如果无法估算出言语特征在相关人群中的分布情况，那么就不能定量地估算法庭语音比较的似然率（在合理的自信度和自信区间内）。但是，同理，利用同一套语料，如果没有这样的估算，也同样无法确定英国体系中的“特殊性”问题，而如何解决非一致性的问题也还不清楚。

也许，英国立场声明的作者和签名者会说：特殊性可以由专家根据他们的自身经验来评判，如果是这样，似然率估算也可以照此进行：专家评判为“格外特殊的”特征值可以用人群中很低的概率来表示（似然率的分母）。

在第六部分，英国立场声明做出结论说：

> “然而，我们认为，由于缺少人口数据以及定义相关人群的问题，因此，在当前条件下还不能将这种方法进行量化应用。鉴于目前存在的一些困难，我们倡导使用上文第四部分中提出的评判体系。” (p. 142, §6)

我们必须设定“当前条件”应该理解为“英国的当前条件”，尽管这一点并没有被完全澄清。在英国当前相当复杂的语言环境下，要确定罪犯或嫌疑人语音样本的口音，目前还没有或几乎没有关于声学特征在那种口音中的分布信息资料。在这种情况下，既不能量化似然

率也不能进行特殊性评估，除非语音专家去人群中收集样本。然而，在澳大利亚和西班牙，法庭科学家已经收集到了分布数据，法庭也已经接受了以似然率形式提供的法庭语音比较报告。这应该归功于澳大利亚的口音要比英国境内的口音少得多这一事实。

就英国的情况而言，问题仍然是：如果没有合适的言语录音数据库，法庭专家就不得不去收集，然后再进行分析。利用全自动系统进行分析，相对比较便宜，但是对于传统的声学-语音学方法，则需要投入相当大的人力去收集和分析参考数据。我们注意到，英国已经着手开展收集数据库和测量特征等进一步的工作了。但是，看起来，短期内英国还不能将证据力度的量化估算应用到一些或者许多案件中。

如果短期的解决方法是对证据力度提供定性的估计，那么，与英国的立场声明相反，采用似然率方法进行评估会比使用英国体系更好一些。Robertson、Vignaux和 Jessen建议，如果没有人群分布数据，可以采用这样的方法：

> “要评价似然率，不必非要有每个概率的精确数字。证据的价值取决于这些数字的比率，因此，如果我们认为证据支持一种假设是支持另一种假设的10倍，那似然率就是10，不管分子和分母的精确值是多少。我们往往能够根据我们的知识和经验大体估算出这些比率。” Robertson and Vignaux (1995: 21)

> “即使在没有人口统计而无法进行量化的地方，也可以采用贝叶斯方法作为概念体系为语音比较分析提供逻辑支撑。” (Jessen 2008:13)

毫无疑问，英国的语音检验人员都是很有资质的，具有丰富的法庭语音比较经验，包括法庭方面的经验和许多特征分析方面的经验。因此，他们对于这些特征的话者自身变异以及话者之间差异的大小都应该有很好的直觉。他们能够以以下方式提供专家意见：“根据我的经验，我认为，根据我所列出的罪犯和嫌疑人语音样本的差异情况，二者来源于同一话者的可能性要比来源于不同话者的可能性大得多。”（或者作必要的修正）。像这样，对竞争假设下的证据概率的定性陈述肯定对法庭是有意义的，这也和我们讨论的将似然率体系作为证据力度评价的正确逻辑体系是一致的。然而，我们同意Lucy的观点，即定性的陈述不是替代量化似然率（和它的置信区间和置信限度）的很好选择。

> “用这样的估计来评估重大的犯罪审判中的关键证据既不理想，也不合意。”(Lucy 2005: 137)

4 总结和结论

这篇回应批评了英国立场声明中提出的关于法庭话者比较的建议。值得表扬的是，英国立场声明的作者和签名者提出了英国的法庭话者比较需要变革，并且提出了变革的方法。特别是，我们看到了他们声明中积极的一面，即避免传统的关于证据的假设概率结论的表述方法所带来的逻辑问题，并且考虑了同时比较语音样本的相似性和典型性。我们也赞同使用“比较”这一术语，并以此来代替辨认、确认和识别等术语。

我们明确了英国提出的建议中的三个根本缺陷。首先，将语音特征作为离散的和不变的特征来对待，就象DNA一样，其结果是很难将该方法付诸实现。其次，我们指出了文章前后的矛盾之处，即先是禁止后又允许使用证据的假设概率，承认缺少支持信息却又使用特殊性表述。第三，缺少关于一致性和特殊性的评价方法并没有帮助事实裁定者对其进行解释。

假定英国立场声明原则上认可似然率方法，我们尽可能地实事求是地给予批评，尽量避免脱离似然率体系进行批评。但是，两阶段评价不是似然率，使用似然率方法才是法庭比较证据评价的现代思想。所以，我们反对英国体系在声明的前言中宣称的：“在概念水平上，与当前提供DNA证据时使用的方法一致(p. 138)。”的观点。最后，我们还要指出的是，英国立场声明并没有达到它的目标。但是，值得称赞的是，“它通过现代思想将法庭语音比较和其它法庭科学结合起来。” (p. 137)

看到英国提出的分析体系应用于研究和案件将是一件有趣的事。当然，我们还是鼓励世界范围内的所有法庭语音比较研究人员和检验人员尽快采用量化的似然率方法为标准（虽然有一些法庭言语特征只能用定性方法来评价）。如果英国现在就投入力量获取定量化数据，为基于似然率方法的法庭语音比较提供支撑，那么，我们希望在英国这不久就能成为现实。

参考文献

Aitken, C. G. G. and Stoney, D. A. (1991) *The Use of Statistics in Forensic Science*. Chichester, UK: Ellis Horwood.

Aitken, C. G. G. and Taroni, F. (2004) *Statistics and the Evaluation of Forensic Evidence for Forensic Scientist* (2nd ed.). Chichester, UK: Wiley.

Aitken, C.G.G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53(4):109-122.

Balding, D. J. (2005) *Weight-of-evidence for Forensic DNA Profiles*. Chichester, UK: Wiley.

Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370–418.

Champod, C. and Meuwly, D. (2000) The inference of identity in forensic speaker recognition. *Speech Communication*, 31: 193–203.

Clermont, F., French, J. P., Harrison, P., and Simpson, S. (2008) Population data for English Spoken in England. Paper presented at the 17th meeting of the International Association for Forensic Phonetics and Acoustics, Lausanne, Switzerland, July 2008.

Donnelly, P. (2005) Appealing statistics. *Significance*, 2(1): 46–48.

Evett, I. W. (1977) The interpretation of refractive index measures. *Forensic Science International*, 9: 209–217.

Evett, I. W. (1991) Interpretation: A personal odyssey. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 9–22. Chichester, UK: Ellis Horwood.

Evett, I. W. (1998) Towards a uniform framework for reporting opinions in forensic science case-work. *Science & Justice*, 38(3): 198–202.

Evett, I. W., and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sunderland, MA: Sinauer Associates.

Foreman, L. A., Champod, C., Evett, I. W., Lambert, J. A., and Pope, S. (2003) Interpreting DNA evidence: A review. *International Statistics Journal*, 71: 473–473

French, J. P. & Harrison, P. (2007) Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech Language and the Law*, 14(1): 137–144

Friedman, R.D. (1996) Assessing evidence. *Michigan Law Review*, 94: 1810–1838.

González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., and Ortega-García, J. (2006) Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20(2-3): 331–355.

González-Rodríguez, J., Rose, P., Ramos, D., Torre, D., and Ortega-García, J. (2007) Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7): 2104–2115.

Good, I. J. (1991) Weight of evidence and the Bayesian likelihood ratio. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 85–106. Chichester, UK: Ellis Horwood.

Haigh, J. (2005) Review of statistics and the evaluation of evidence for forensic scientists. *Significance*, March: 40.

Hodgson, D. (2002) A Lawyer looks at Bayes' Theorem. *The Australian Law Journal*, 76: 109–118.

Hudson, T., de Jong, G., McDougall, K., Harrison, P., Nolan, F. (2007) F0 statistics for 100 young male speakers of standard Southern British English. In J. Trouvain and W. J. Barry (eds.) *Proceedings of the 16th International Congress of Phonetic Sciences* 1809–1811.

Jessen, M. (2008) Forensic phonetics. *Language and Linguistics Compass*, 2(4): 671–711.

Lindley, D. V. (1991) Probability. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 27–50. Chichester, UK: Ellis Horwood.

Lucy, D. (2005) *Introduction to Statistics for Forensic Scientists*. Chichester, UK: John Wiley.

Nolan, F. (1983). *The phonetic Bases of Speaker Recognition*. Cambridge, UK: Cambridge University Press.

Nolan, F. (1996) Forensic phonetics. Notes distributed at the two-week course at the 1996 Australian Linguistics Institute, Australian National University, Canberra.

Nolan, F. (1997) Speaker recognition and forensic phonetics. In W. J. Hardcastle and J. Laver (eds) *The Handbook of Phonetic Sciences* 744–767. Oxford, UK: Blackwell.

Nolan F., McDougall, K de Jong, G., and Hudson, T. (2006) A Forensic Phonetic Study of 'Dynamic' Sources of Variability in Speech: The DyViS Project. In Warren & Watson (eds.) *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* 13–18.

Pigeon, S., Druyts, P., and Verlinde, P. (2000) Applying logistic regression to the fusion of the NIST '99 1-speaker submissions. *Digital Signal Processing*, 10: 237–248.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10: 19–24.

Robertson, B. and Vignaux, G.A. (1995) *Interpreting Evidence*. Chichester, UK: Wiley.

Rose P (2002) *Forensic Speaker Identification*. London & New York: Taylor and Francis.

Rose P (2003) *The Technical Comparison of Forensic Voice Samples*. Issue 99, *Expert Evidence*. Freckelton I, Selby H, (series eds.). Sydney, Australia: Thomson Lawbook Company.

Rose (2005) Forensic Speaker Recognition at the beginning of the Twenty-First century. An overview and a demonstration. *Australian Journal of Forensic Sciences*, 37(2): 49–71.

Saks, M. J., and Koehler, J. J. (2005) The coming paradigm shift in forensic identification science. *Science*, 309: 892–895.

Thompson, W. C., and Schumann, E. L. (1987) Interpretation of statistical evidence in criminal trials. The prosecutor's fallacy and the defence attorney's fallacy. *Law and Human Behaviour*, 11: 167-187.