# 文本相关的说话人确认:分类器，数据库和RSR2015

计111（10111939）陈楚楠

摘要：已经被新加坡Infocomm研究所的人类语言技术部门收集和发布的RSR2015数据库，旨在评估文本相关说话人识别在不同时间和词汇的限制。依据多种口音的多样性的平衡，记录说英语的人的语音。有超过151小时的语音数据通过移动设备记录。说话人包括300名17至42岁的参与者(143名女性和157名男性)。使得RSR2015数据库成为最大的公开的针对文本相关确认的数据库。我们为数据库的三个部分的每个部分都提供评估协议，连同两个说话人确认系统的成果：HiLAM系统，基于三层声架构的i-vector/PLDA系统。因此我们提供一个相关的评估方案，以及参考RSR2015数据库性能研究社区的演示方案。HiLAM在大多数场景下比使用了最新技术的i-vector性能好。

关键词：文本相关，RSR2015数据库，说话人确认

## 1介绍

说话人确认过程是指通过比较两个语音样本来接受或拒绝身份，这两个语音样本:一个用作参考的身份，另一个是在测试期间收集的其他的人的语音。在这个通用的定义下，说话人是可以无偿提供的任何没有时间，质量，记录条件和词汇的内容的约束的语音样本。说话人确认的性能受到许多可能的说话方式可变性的影响，其中在词汇内容和通道的变化是最不利的。

普遍认为，当测试话语与给定的材料相同，尤其是当其内容比较小时，能实现更好的精度。在这方面，有两种方法能明确有效地解决词汇变化的问题。第一种方法包括进行后验分析抵消训练和测试之间的话语中词汇不匹配的语音样本，而第二个方法考虑对于合作的说话人，词汇的变化可以很容易地降低。首先，尽管第一个方法较高的灵活性，它会受到来自两方面的缺点。一方面，验证任务的词法分析将增加计算成本。另一方面，词法的抵消可能会受到大量的词汇不匹配，因为它不可能保证初始的训练和测试词汇一样。第二种方法可以要求合作的说话人在训练和测试阶段说出一个预先定义的句子或短语。这个过程称为文本相关说话人确认与没有约束输入的文本无关说话人确认相反。换句话说，文本相关的说话人确认可以定义为一个说话人确认任务，其中在测试阶段使用的词汇发音是说话人在训练时的一个子集。

相比于无法控制的环境因素所导致的通道变化，如果我们可以假定说话人是配合的，词汇的变化是相对更简单的。依据文本相关假设，通过要求测试话语的词汇内容匹配训练材料，我们在训练和测试阶段可以达到更高的精度和更短的时间。因此，文本相关说话人确认对人体工程学的精度要求高的，记录持续时间短的商业应用非常适合。

在缺乏词汇约束的训练/测试话语时，文本无关技术涵盖广泛，应用的范围包括，法医确认、说话人集群和说话人挖掘。此外，在上下文有大量的数据可用时，研究文本无关任务是国家标准与技术研究所(NIST)强烈支持举办的国际基准事件。由于这些原因，大量科学论坛关注文本无关说话人确认，而轻视了文本相关的说话人确认的商业潜力。无论如何，文本相关说话人识别可以视为文本无关的说话人确认的一个补充。当确认话语有类似的持续时间和词典时，可以弥补当前的不足并且获得更灵活的性能。从历史观点上来看，本文支持文本相关的说话人确认作为文本无关的说话人确认的子情况的一系列的改进。

最近通过集中加强开发和使用数据，在文本无关的说话人确认的精度和鲁棒性方面的有了突破。这些改进一直由几十年来一直提供越来越具有挑战性数据的NIST和LDC强烈支持着。虽然文本无关说话人确认正在被越来越多地训练数据健壮的系统，而文本相关说话人确认苦于缺乏研究数据。这就导致在文本相关这种子情况下，适应现有的方法存在不少的困难。例如，在《使用文本相关说话人节点因素确认》这篇论文发表9年之后，没有与文本相关的说话人确认的论文被发表。这可以理解为开发这样的系统缺乏支持文本相关任务的数据库。

数据的缺乏在不同方面上影响着文本相关的说话人识别。现有数据库的约束不允许适当的来自词汇变化成为选择说话人的条件。此外，在第二章的现有的数据库普遍是性别不平衡的，然而在大多数时自动系统的表现是要求能区分不同的性别。最后，自动确认系统需要大量的在统计学上有意义的措施来改进。

在文本相关的说话人确认下，演讲的词汇内容数据尤为重要，有很多方法可以限制训练和测试话语的词汇。我们可能需要解决词汇约束各级如音素、音节、单词或句子，因此有不同的确认方案。多项研究表明，保留确认话语中的词汇序列可能导致错误率相对下降50%。因此，从应用程序的观点选择一个特定的方案是至关重要的，因为它会强烈地影响精度。然而，很少有比较研究对不同词汇的影响进行约束，一定程度上是由于缺乏一个可以支持公平比较研究的数据库。

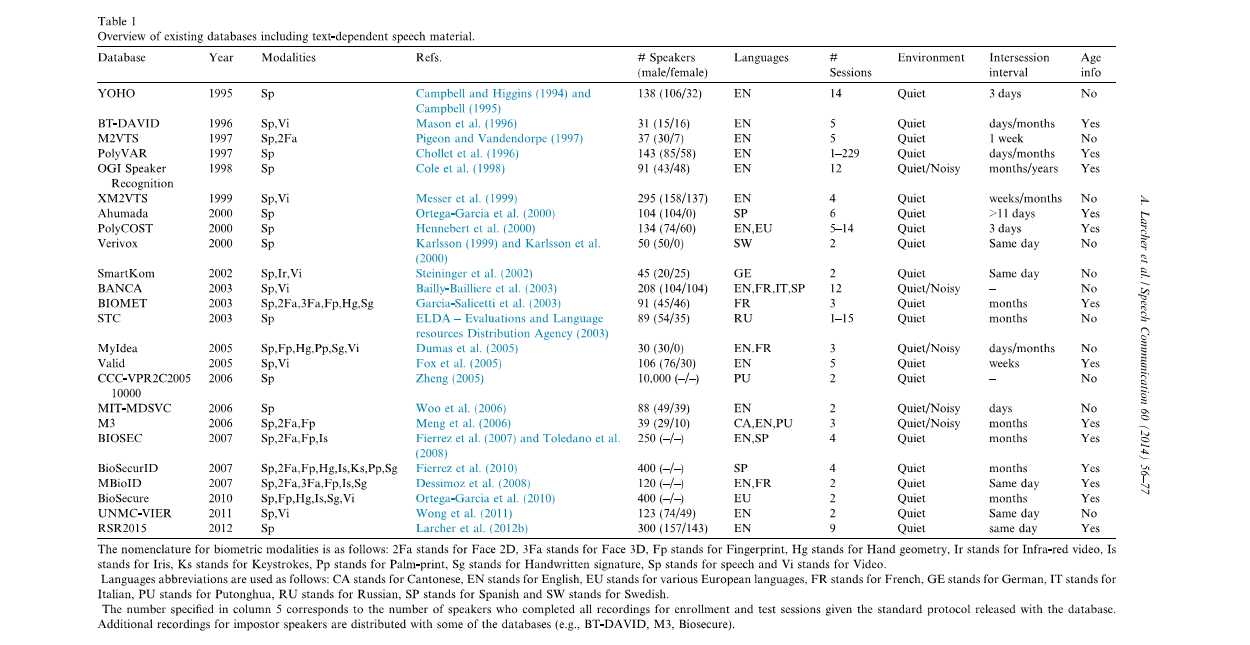
在本文中，我们展示了在的Human Language Technology Department发布的RSR2015数据库，以解决的一些现有的语料库的局限性的问题。这是部署在智能家居的HOME2015计划下的鲁棒说话人识别记录的一部分，也是RSR2015数据库的名称的由来。RSR2015数据库的目的是支持文本相关的说话人识别，并允许在确认系统上研究存在不同词汇和时间的约束。RSR2015数据库是最大的公开的可用的文本相关的在说话人和词汇的变化的说话人确认数据库，涉及143名女性和157名男性演讲者总计151小时的录音。数据库被安排成三部分，以适用不同的场景。所有的三个部分都被在相似的条件下记录，以保证任务之间有一个公平的比较。获得RSR2015数据库后可以在六个包括智能手机和平板电脑的不同移动设备上商业化使用。RSR2015数据库的第一部分被Larcher等人描述。

剩余的这篇文章中，我们首先概述现有的其它的文本相关的说话人确认数据库。我们通过他们提供的描述性的文献中的主要特点、优点和缺点，总结了23个数据库。RSR2015将在第三章进行详细的描述。在下面的章节中，我们提出了对RSR2015数据库和其它系统的切合实际的评价方案和绩效措施，以做出公平的比较。在第四章中，我们给出了一个用于文本相关的说话人识别的概况分类，在描述两个先进的评估RSR2015数据库的系统之前。第五章描述了详细的方案和RSR2015数据库的三个部分分别的两个系统的性能表现。第六章提供实用信息关于如何获得这个数据库。在第七章我们还将讨论一些关于文本相关的说话人确认的研究方向和观点。

## 2文本相关说话人识别数据库

在本节中，我们将展示目前语音数据库提供的文本相关的说话人验证评价的发展概况。尽管给定的数据库(表1)可能不是最详尽的列表，但它尽可能地包含了我们所知的尽可能多的文献。说话人识别技术互补的信息资源可以在Campbell & Reynolds和一项Faundez-Zanuy给出的多模式生物数据库概况中找到。值得一提的是，有一些文献报道的结果来自不公开可用的数据库。

我们的目的是提供一些和RSR2015数据库特性有关的而不是给出一个详尽的关于现有的数据库的描述。事实上，不同的文本相关的说话人确认数据库被设计用于不同的目的，另外它的多样方案使得其很难有一个公平的整体的比较。在剩下的这一章中，我们讨论一些现有的文本相关的说话人确认的数据库的主要特征。

表1 现有的包含文本相关语音材料的数据库综述

### 2.1　人口统计

在评估说话人确认系统的性能时，人口统计是至关重要的。在说话人确认中，开发和评估自动系统是由现有的语料库驱动的，它的人口记录的数据集是由精心挑选而来的。特定的应用程序，选择人口尽可能代表的目标人群，而数据库为了通用的研究目的往往设计成覆盖最大的可能的人口。在剩余的这篇文章中，我们考虑的是人口的性别和年龄，它们通常被视为影响说话人确认引擎的两个主要条件。出于这个原因，说话人的集合应该代表目标人群的性别和年龄。此外，人群需要足够大，才能在性能上通过自动说话人确认系统的大量的试验改善，以确保结果在统计上是有意义的。

在实践中，人口的规模，还要考虑技术和经济问题等因素。有趣的是，数据库的一大部分是文本相关的语音材料，是多模数据库，也就是说在表1中列出的24个数据库有14个是多模的。收集多模式的数据库的好处是双重的。首先它允许研究比较和组合形式的身份确认。其次，数据池成本和复杂性的负载，也与收集生物样本相同。从表1可以看到，在过去的20年中，科学界通过巨大的努力已经收集了维持生物识别技术的研究的数据。然而，这些数据库中的说话人相关的数量仍然有限，只有7/24个表中的条目数超过200。

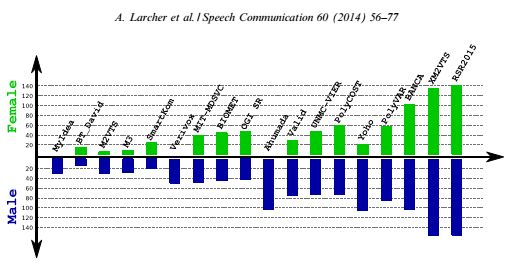
另一个约束是可以从到图1观察到的性别的不平衡。若有19个数据库的性别信息是可以检索的，则有8个的性别比达到了45%故是性别平衡的，而有7个包含少于30%的女说话人不是性别平衡的。这种性别的不均衡在区别说话人的性别时特别损害说话人确认系统的性能。此外，年龄并不总是可用的信息(至少在表1中列出)。当年龄差距很小时想要找出说话人之间的差别已被证明是更加困难。

图1 比较每个公开的文本相关说话人识别数据库的每个性别的人数量。只有含有性别信息的数据库是可用的，且按照说话人总数递增排序。

### 2.2　词汇的变化

说话人确认系统的性能是强烈依赖于作为输入提供的语言材料。例如，许多研究已经进行了估计说话持续时间的影响的研究。其他作品表明，分歧率取决于用于注册和测试的演讲内容，主要的结论有，在固定的时间，相同的话语对说话人确认相当有用。在文本相关的说话人确认中，注册和测试话语都是固定的，词汇的内容尤其重要，因为它会影响系统的精度。因此，选择的词汇内容时，应研究其对文本相关说话人确认系统的影响。

#### 2.2.1说话人确认的主流方案

在过去的20年中，NIST所提供的大型数据库及其相关方案已成为文本相关说话人确认技术的事实上的标准评价。在文本相关说话人确认的情况下，没有导致整个系统的比较更困难的标准和增加文献报道的方案的标准存在。然而，两个主要流反映在现有的数据库。在 Yoho，M2VTS，Verivox 和 Biosec，词法的训练和测试内容被强烈限制只能使用位数，而SmartKom，STC，CCCVPR2c2005-10000，或MIT-MDSVC，则允许更广泛的词汇覆盖使用固定短语的。这两种类型的协议都被被一些数据库一起使用如Polyvar，OGI Speaker verification，XM2VTS，Ahumada，PolyCost，BANCA，BioMet， MyIdea，Valid，M3，BiosecureID，MBioID，BioSecure，或提供不同的数字组串的固定短语的UNMC-VIER。大多数情况下，词汇的变化范围是在有限的几个固定的句子和固定数字字符串。例如，在表1列出的24个列表中有10个数据库仅含少于10个不同的句子。

为了增加词汇覆盖率，一些数据库如PolyCost，Banca，MyIdea或MIT-SDSVC包括词汇内容不同的说话人。试验中的冒充者说出目标说话人所使用的要求每个主题定一些其他学科的内容的文本。在这种方案下，说话人交互的冒充者试验的通过可能性被大大限制了，因为冒充者和目标说话人不都说同样的演讲内容的情况。

#### 2.2.2 文本相关的说话人确认的语音

语言词汇内容一直被各种语言所限制，如法语、德语、意大利语、汉语、俄语、西班牙语或瑞典语，近年来社区一直在努力地提供资源。一部分的数据库得益于由于欧盟内的协作努力，也包括多语言内容。然而，在表1中列出的24个数据库的10只包含英语演讲，而另外6个包括其他语言，英语仅作为补充。因为英语是除了母语最容易的，另外英语由于历史原因可以作为国际标准，所以英语在现有方案中最全面。

### 2.3 会话的多样性

匹配训练和测试之间的词汇内容，可以使得不匹配的情况大大减少，使说话人确认任务更容易在短期内处理。然而，其他我们称之为会话变化的因素，如通道失配，环境噪声或说话人内在的可变性，还影响着说话人确认系统的性能。在本文的剩余部分，会话这一术语是由一个或多个不同的录音环境等因素，录音设备或时间的录音来区分的。

由于复杂性和数据采集的成本，尤其是对多式联运方式的全集，大多数数据库记录使用相同的麦克风和控制环境，以强烈限制跨会话通道和噪声变化。(例如Yoho，BT-David，M2VTS，XM2VTS，Verivox，SmartKom，Biomet，STC，Biosec)其他数据库通过提供的录音室外音箱等各种环境关注不利条件，例如在街上，在公共区域如建筑大厅或餐厅。(例如 BANCA，MyIdea，MIT-MDSVC or M3)这些数据库在表1的第八列被用“噪声”标注，用来表示相比于其他数据库没有明确解决环境的不匹配。最后，有些数据库包括在不同的设备，但不强加任何背景噪音或环境因素记录下的，与说话人的显式通道不匹配。(例如 PolyCOST或PolyVAR)

文本相关的数据库的会话数量往往有限，由于成本和记录的次数与时间成正比，另外还必须动员说话人。在表1列出的24个数据库中，有16个数据库的每个说话人的会话少于5个。大量的数据库，特别注意两次之间的时间间隔内的记录，最大化说话人在众所周知的生物衰老的影响。然而，(Lei & Hansen， 2009; Lawson等人，2009; Kelly 和Harte，2011; Kelly 等人，2012)表明，文本无关说话人确认的情况下，老化效应只在若干年后才变得有意义，因此只有OGI说话人识别数据库覆盖到。相反，其他文献认为一段时间后或几个月后文本相关说话人确认的案例出现显著恶化。这些研究表明，老化现象还不是很清楚，可能以不同的方式影响文本相关和文本无关说话人确认。

## 3 RSR2015 数据库

RSR2015数据库，为社区提供一个足够大的数据集的性别平衡的说话人记录。它由多个手机和平板电脑记录，在9个会话中有300个说话人。196844个文件记录包含了151小时30分钟的音频信号。

为了让说话人的确认系统在不同词汇的约束下公平比较，一直特别注意到词汇内容。因此，记录的300人分为三个部分，每个致力于特定的任务，涉及不同的词汇和时间约束。第一部分致力于说话人使用固定短语句确认。第二部分致力于说话人引导的命令控制。第三部分致力于说话人使用随机提示的数字字符串确认。为了允许用例之间的公平的比较，三个部分已经在类似的说话人和通道条件下记录。

### 3.1 人口统计学

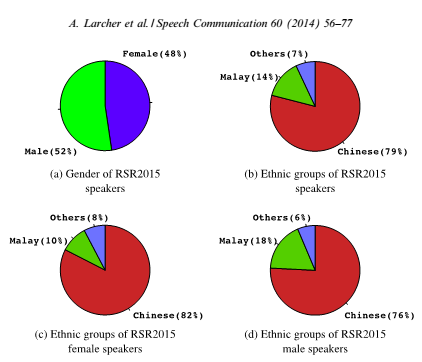
RSR2015最大的公开的文本相关数据库之一，有300个说话人记录。据我们所知，包括文本相关说话人的数据库最高级别的是BioSecurID和词汇内容小于RSR2015数据库的CCCVPR2C2005-10000。

图2 RSR2015数据库中300个说话人的性别和种族统计。特别注意这已经被移植成平衡性别(女:143男:143)和反映新加坡人口的种族起源(中国:237,马来语:42和其他人:21)。

为了代表新加坡人口，RSR2015数据库记录的300人是根据新加坡的种族和性别分配的。性别平衡的结果是数据库中的300人中有143人是女性。(图2(a))另外，237人是华裔，42人是马来裔，而剩下的21从其他不同的民族组。(图2(b)-(d))。

RSR2015数据库包括从17岁到42岁的说话人 (图3) 。鉴于RSR2015数据库的样本量有限，若扩大年龄创建一个稀疏分布的不同年龄的说话人集合，可能会人为地促进说话人确认的任务。这被Doddington(2012)确认过，说话人确认任务难度的增加与说话人之间的年龄差距是有限的。因此说话人在一个有限的年龄段可增加说话人确认的要求。

### 3.2 获得协议

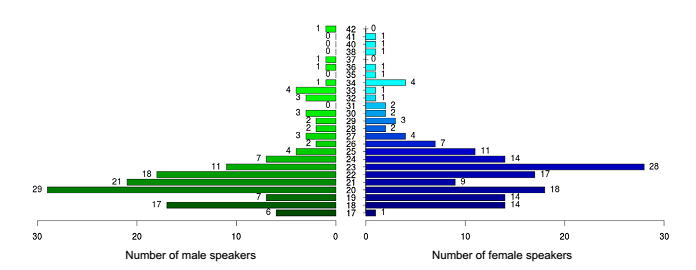
RSR2015数据库的的9个会话记录了室内的一个典型的办公环境。每个主题的录音过程都在一天完成，所以RSR2015数据库不包括衰老变化。相比与会话变化，在3年内的年龄变化已被Lawson等人证明可以忽略不计。

图3 RSR2015数据库的说话人的年龄和性别金字塔图。

市场上可用的6个移动设备(5个智能手机和平板电脑)被用来记录。3个便携式设备(标记为A、B和C)被分配给每个主题。每个主题的9个会话使用三个设备的顺序会被记录下来：{A，B，C，A，B，C，A，B，C}，和元数据信息一起提供。并用Android应用程序实现一个对话管理器，来管理记录。这个应用程序使用设备的触屏功能提示文本内容。使用即按即说-特性来允许用户启动录制和停止后阅读提示。使用便携设备的目的是舒适和自由，在会话内或者跨会话音质也可以有明显变化。

音频信号被6个便携设备通过内部麦克风处理，记录成原始PCM格式，采样率为16kHz，采样位数为16bit。之后在添加一个包括关于设备的信息，语言，样品的数量，采样率和数据格式的SPHERE头。

### 3.3 3种不同的文本相关任务

可以用不同的场景来约束训练和测试的话语的词汇内容。尽管不同的词汇和各种场景所需的时间约束各种各样，现有的数据库很少包括允许比较系统的场景的数据。表1中列出的24个数据库中只有8个包括可以用来研究会强烈影响确认系统的性能的协同发音效应材料。为了能在不同的场景系统中比较，RSR2015数据库设计的三个部分分别用不同的词汇约束，并同时保持相同的记录情况。在下面描述的三个部分的所有300个说话人，在一个给定的会话中发音的词汇内容都是相同的。在每个会话中，每个说话人要求对每个句子发音一次。

RSR2015数据库的第一部分关注文本相关说话人确认任务，说话人通过对固定短语的发音进行身份确认。在9个会话中，说话人说出30个来自涵盖所有英语音素的固定的句子的TIMIT数据库。跨说话人，会话记录和句子的平均时间是3.20秒，每个句子的平均持续时间从2.73到3.65秒变化。请注意，这些句子已选定具有接近的持续时间来评估不同词汇的影响。在第5章中描述的基于能量的语音活动检测(SAD)之后，整个句子平均名义演讲时间是1.25秒。(不同的句子从1.01到1.59秒变化)RSR2015数据库的整个第一部分包含72小时的录音。(在SAD之后有28小时15分的名义语音)

RSR2015数据库的第二部分着重于说话人引导的命令控制任务，说话人通过固定命令的发音来进行说话人确认和控制家用电器。在9个会话中，说话人说出30个简短的命令用来控制位于新加坡Fusionopolis的功能齐全的180平方米智能家居原型StarHome中的家用电器。跨说话人，会话记录和句子的平均时间是1.99秒，每个句子的平均持续时间从1.66到2.46秒变化。在SAD之后，整个句子平均名义演讲时间是0.63秒。(不同的句子从0.44到0.99秒变化)RSR2015数据库的整个第二部分包含44小时53分钟的录音。(在SAD之后有14小时12分的名义语音)

RSR2015数据库的第三部分着重于提示用随机数字序列完成文本相关的说话人确认任务。在9个会话中的任意一个，说话人说出3个由10个数字组成的序列和10个由5个数字组成的序列。在跨会话时数字序列是不同的，是根据说话人身份生成的，冒充着也需要正确的发言序列。用于训练和测试的语音材料被约束在十个数字(0-1-2-3-4-5-6-7-8-9)但训练和测试时时，上下文之间的每个数字的左右都是不同的，以评估协同发音效果。

对于10个数字组成的序列，跨说话人，会话记录和句子的平均时间是5.19秒。在以应用能源为基础的语音活动检测(SAD)之后，整个句子平均名义演讲时间是2.07秒。对于5个数字组成的序列，跨说话人，会话记录和句子的平均时间是3.06秒。在以应用能源为基础的语音活动检测(SAD)之后，整个句子平均名义演讲时间是1.09秒。RSR2015数据库的整个第三部分包含34小时36分钟的录音。(在SAD之后有12小时51分的名义语音)

## 4 文本相关说话人确认分类器

文本相关说话人确认系统想要进行有意义的精准度高的比较往往是非常困难的，因为RSR2015数据库缺乏激励收集的评估标准方案。此外，系统架构可以影响特定用例和其固有的词汇约束。在本节中，我们首先概述用于文本相关说话人确认的现有的分类器。在第5章，我们将描述两个先进的说话人验证系统:基于GMM和HMM的文本相关的系统建模和RSR2015数据库的表现中的i-vector系统。这些分类器的选择是出于当前发言者的代表性确认引擎。事实上，文本相关引擎最近部署在一个大规模的商业应用程序而i-vector系统改编自文本无关说话人确认的主流的先进的引擎。为了建立一个公平的比较，两个系统都使用相同的前端，另外使用两个性能指标描述他们的表现。

### 4.1 现有分类器的概况

图4 对应RSR2015数据库的三个部分在几秒钟内的分布的录音(浅蓝色)和名义上的演讲时间(暗红色)。第三部分,持续时间有10位序列分别和5位序列。（对引用这一图表色彩，读者可以参考这篇文章的网络版）

文本相关说话人确认系统的特异性，即他们必须对说话人特征与内容验证的话语的词汇进行建模。在过去的三十年，两个主要的趋势一直在主导着文本相关说话人确认的领域。

第一个分类，基于动态规划提出了可用演讲的数量是有限的。它们工作在框架层面，提供了一个精确建模的时间结构提供的语音发音，但是缺乏概括能力生成方法。特别是，作为动态编程技术通常提供欧几里得距离而不是语音样本之间的似然概率，因此需要额外的确认说话人的决策机制，它通常是作为概率假设检验领域制定。通过引入距离标准化或多模板方法，已经多次弥补说话人内在的变化。如超音段和源特性的额外的信息，也可以用来增强系统。

第二个分类器，是目前最常见的，是基于隐马尔可夫模型(HMM)。隐马尔可夫模型是更强地适应语音信号的变化，可以利用一个更大的数量或训练数据。此外，取得进展的领域使得文本无关的说话人识别和语音识别受益。在实践中，文本相关的说话人确认面临不同的使用情况，每一个都有一个独特的建模和运行时的要求。依照HMM模型，模型的粒度可以根据代表的演讲话语的时间结构订制。基于手机系统提供优秀的粒度可用于任何词汇内容在HMM建模词组或整个话语，其粒度越小，越受制于有限的词汇。研究还进行了改善这些模型的鲁棒性的信道和扬声器的变化。在Chatzis和varvarigou(2007)中，隐马尔可夫模型的状态的高斯分布被Student-t分布代替，面对噪音时有更强大的鲁棒性。在Aronowitz(2012)，作者采用的支持向量机的概念与扰动属性投影(NAP)，也是使用隐马尔可夫模型。尽管这种方法的性能良好，但它是有限的情况下，系统的所有用户共享相同的密码，因为NAP矩阵训练所需的数据量。

在其他的文献作品提出了利用人工神经网络或基于表示频谱使用的语音的时间结构模型。不同的建模方法，最终可以结合以弥补各自的弱点。

最后，文本无关的说话人确认系统也从一些系统中找到灵感。在经典的基于GMM/UBM或更近的 i-vector表示的体系结构已被证明是在采用文本相关的说话人验证所需的词汇信息有优势。这些系统的成功都有局限性，因为他们没有明确地使用好言语的时空结构的优点。然而它有可能将这些方法与语音识别引擎结合，有效地在一个额外的计算成本交换空间中确认说话人和语音。

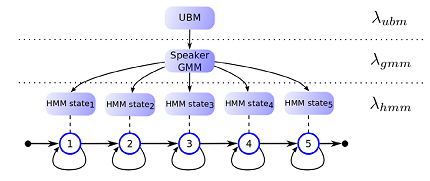
从贝叶斯观点，基于生成的方法系统是优于动态规划的，可以产生似然比得分，在决定的时候容易校正和解释。在下面，我们提出了两个最先进的文本相关的说话人确认系统：HiLAM说话者验证引擎基于GMM和HMM建模和i-vector系统，这两者都产生自然的似然比。

HiLAM的鲁棒性已经受到了广泛的测试，因为其一直在商业部署。i-vector系统是来自最先进的文本无关的说话人识别引擎

### 4.2 文本相关系统HiLAM

#### 4.2.1 话语的建模

分层的多层声学模型(HiLAM)是一个已被Lee等人和Larcher等人描述的文本相关的说话人确认引擎。图5是经典的GMM/UBM方法的一个扩展描述。在这种架构中所有的发射概率的混合高斯分布(GMM)共享相同的方差和权值参数。前两层类似于标准GMM/UBM在上层模型一般语音的声学空间的UBM。中间层是文本无关的说话人模型的一个经典的UBM的最大的后验适应。底层的能力取决于隐马尔科夫模型(HMM)从左到右利用密码短语的具体时间结构。每个HMM状态发射概率密度函数是来自中间层的说话人依赖GMM。这些GMM改编自跟随MAP标准的文本无关的说话人模型。适于只有平均参数，它不同于Larcher等人最初提出的权重参数的调整。使用连续密度隐马尔可夫模型(CDHMM)，这基本上取代了原工作中的计算费用高的半连续隐马尔可夫模型(SCHMM)，以提供更高的精度。

HiLAM的训练和Larcher等人原来的描述类似。首先训练性别独立的UBM模型的声学空间。改编自所有数据明显的目标说话人的UBM的文本无关的说话人模型。最后一个迭代训练进行训练的第三层HMM模型。为了对每一个密码初始化HMM，话语切成相同长度的S段。HMM的每个状态适于使用GMM的中间层的相应的。然后利用HMM改编一个新的分割。这个过程中使用到Viterbi算法。这个迭代过程直到Viterbi路径收敛。状态数的选择是根据经验的。HMM的从左至右的转换概率设置成相等的。在测试过程中，给定一个语音序列X，一个文本相关的评分，，计算公式

是说话人文本相关的HMM对齐的维特比译码，和X在UBM的似然比，的对数比。当建模的句子从第一部分开始时，每个半连续HMM的状态数在依照经验设置为5，它被设置为3时，较短的命令从第二部分建模。

图5分层的多层声学模型

#### 4.2.2 数字的建模

一种改进的HiLAM版本已经发展到可以处理随机提示的数字。该结构的两层类似原始模型。在训练时，每个说话人对几个10位英文数字发音。目标说话人的录音是采用说话人GMM从中间层采用最大后验概率MAP准则自动分割成含有十个高斯百位数的训练集。请注意，训练的话语的分割是通过使用一个先进的语音识别系统的状态，因此没有迭代适应进行训练的HMM组成。在测试过程中，给定一个随机提示序列的N个数字，从左到右的HMM组成与相应的N个数字模型。确认得分计算式是根据的使用Viterbi对齐得到的试验段在HMM的拟然性。

### 4.3 标准i-vector系统

i-vector的模式是基于假设的可变持续时间的语音段，可以表示为固定长度的向量的i-vector，在低维空间称为完全变异的空间。利用完全变异的空间的维数低的优点，许多分类技术也被用于执行不同的任务，如说话人和语言识别或说话人分类。因为i-vector可在语音段保持不同类型的变化，如说话人和词汇内容，最近的工作表明i-vector可用于文本相关的说话人识别的任务。

#### 4.3.1 i-vector提取和标准化

一个语音段上的总变异的空间投影可以视为一个根据GMM方法降低了通道和扬声器的超向量级联概率压缩过程的维数的线性高斯混合模型。超级矢量，投射到总变异的空间根据生成的方程：

其中是一个说话人和通道独立的超向量，是一个因子载荷矩阵低秩和是一个服从标准正态分布的随机向量。一个i-vector x是估计给定一个语音的最大后验概率的点。关于i-vector提取过程的更多细节可以在Dehak等人和Martinez等人著作中找到。

大部分采用i-vector分类技术都假定他们遵循高斯分布，但在现实中不是这样的。几个标准化的算法已被提出，根据高斯假设修改i-vector分布。球形滋扰标准化，SphNorm，已经在在概率线性判别分析相结合PLDA中显示出良好的性能。

SphNorm是一个参数估计在一个大的开发集i-vector的迭代过程的。对于每一次迭代，n是和的平均值是开发集计算出的类内协方差。所有从发展组的i-vector **x**，在之后根据以下算法归一化。

基于球形滋扰标准化算法的i-vector标准化

给定一个测试向量**x**，从n=1开始迭代：

来自测试集的i-vectors将在之后规范化后成相同的的转换。

#### 4.3.2 说话人的概率线性建模

判别式分析

在Prince和Elder的介绍中，PLDA是一个生成模型，假定观测值是在四组件中的会话j的说话人i。

其中是i-vector的分布的平均值，**F**和**G**是低维矩阵，其列向量构成了两个子空间的秩，分别包含说话人和会话的变异。因此，和是潜在的和子空间相关的变量。是一个标准的分布式的添加噪声，来自协方差矩阵和条件和下面给出的先验密度：

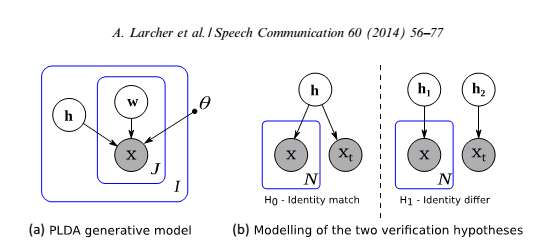
PLDA的图形模型如图6所示，用于实施如Jiang等人和Lee等人的工作。

图6 (a)PLDA生成模型的图形模型。对每个说话人I, J i-vectors x 在总变化空间中观察。PLDA模型被用一个集合描述。(b)原生PLDA框架的两个验证假说的图形模型。

#### 4.3.3. i-Vector 配置

对于这项工作，i-vector的配置已根据对开发数据优化的经验选定。一个从NIST SRE 2004，2005，2006数据库中的12706次会话中的性别无关的2048分布的对角协方差矩阵的UBM模型已经被训练。性别无关的秩为400的总变异矩阵在之后在66702次会话从SwitchBoard II 阶段2和3，SwitchBoard Cellular，第一二部分，Fisher English 和 NIST-SRE 2004，2005和2006的数据库用Kenny和Dumouchel描述的迭代的10次EM算法训练。一个性别无关的PLDA模型是从RSR2015数据库的背景集的50名男性和47名女性说话人的26136次会话中估计而来。矩阵F的秩设置为400，矩阵G被设置为零和并且为满。

在我们以前的工作中我们发现，采用一个适当的定义去训练SphNorm和PLDA，能提高i-Vector系统在对文本相关的情况下的说话人验证系统性能。因此，在第一和第二部分的实验中，分类的定义考虑话语的说话人的身份和第三部分每个说话人训练的词汇内容。所有的组件其载体系统已被开源工具包ALIZE使用。

## 7 总结

本文的三个贡献之中，主要是发布了大型语料库RSR2015。RSR2015的数据库，收集和提供了以允许比较在不同的时间和词汇约束的文本相关的说话人确认算法。所有的人在不同的会话重复同样的密码短语，我们相信RSR2015数据库也可以被用来进行反回放分析。RSR2015数据库从300性别平衡的语音中记录151小时语音信号，是一个公开的文本相关的说话人确认最大的语料库。

第二贡献，我们创造的文本相关的任务，是在现有的所有文献中我们所知的语音数据库最大的库存。我们提出的趋势和RSR2015数据库设计的主要特点引导着现有的其他数据库。尽管社区近年来通过巨大的努力产生大量的可用数据库，但我们强调了几个现有的数据库的缺点。大量说话人的需求，代表者们的性别平衡和持续时间和词汇变异，都积极驱动着RSR2015数据库的收集。最近的出版物将RSR2015数据库贡献的资源密集型方法验证文本无关的任务来填补文本相关和文本无关的研究领域之间的差距。在调查这些数据库时，我们制作了一个描述现有的为文本相关的说话人确认的分类器。

本文的第三个贡献是提出了RSR2015数据库的三个部分的评估协议。该协议允许RSR2015数据库中的不同任务的算法的比较。两个系统的性能，给出了一个基线，和另外的不同的协议相比，文本相关系统HiLAM基于GMM和HMM模型，而先进i-vector/PLDA系统基于开源的ALIZE工具集。

实验结果表明，基于GMM/HMM的系统在大多数配置环境下优于i-vector系统。这证实了众所周知的近年来已被广泛研究的i-vector系统在短持续时间上的弱点。在短持续时间的文本相关的说话人验证上下文i-vector的系统行为与以前的文本无关的说话人确认的研究相一致，对男性达到较低的错误率。相反，在HiLAM系统上执行的女性说话人在RSR2015数据库的两部分I和II表现更好。这种情况将在未来的工作中通过对说话人的分布和语音信号中的词汇信息的调查来关注。

对于固定密码短语的情况（部分I和II），我们发现它相比于冒充者播放正确的通行短语的文本记录语音的攻击的情况下更容易拒绝冒充者播放目标说话人对于不同于预期的密码短语的语音。这证实了词汇信息的观测是由短的语音段支配，即使是i-vector的系统的情况下，尽管对话语的时空结构考虑不足。这两个系统的性能是被协同发音不匹配的随机提示数字场景强烈影响的。(RSR2015数据库的第三部分)相比第一部分提供类似语音的持续时间，通过协同发音不匹配引起的退化是相当于或高于He´bert，Kato和Shimizu的报告。i-vector系统中观察的误差增长率展示的方法，没有利用语音信号的时域信息，而是受协同发音的影响。

RSR2015数据库正在记录更具挑战性的条件。这一部分主要包括第一部分的在海上VHF信道传输。