Vocollect®
a business unit of Intermec®

仓库的语音辨识技术选择

2010 年 8 月

Vocollect 白皮书

目录

执行摘要 1
我们想要从仓库中的语音辨识器获得什么？ 1
语音辨识器如何工作 1
为何语音辨识如此困难？ 2
简化问题 3
仓库语音辨识器的多种选择 4
成本影响计算 6
测试结果 7
衡量业绩 8
结论 9

执行摘要

此白皮书回顾了基于电脑的语音辨识器的一些基本要素，包括最大化业绩的挑战。随后我们根据仓库员工采样语音辨识器的需要，对技术的选择进行了讨论，并计算了仓库环境中识别器错误的成本。对于主要设计决策之一，即采用经培训的（“与说者有关”）还是未经培训的（“与说者无关”）的识别器，我们已证明，花费在训练识别器上的时间，很可能在数天之后就会通过改善的业绩得到补偿。

我们想要从仓库中的语音辨识器获得什么？

在深入研究技术细节之前，先让我们看看使用者 / 客户对仓库语音辨识器的目标的看法。在理想状态下，语音辨识器的主要目标在于，“立即正确地理解用户听到的一切，没有不能理解的。”然而，从以下内容中我们可以看出，这种理想状态是不可能达到的，因此我们不得不谈谈如何尽可能地接近理想状态。理解了这种限制性，仓库识别器应该：

- 在各种声音环境下有效地工作，声音环境包括十分安静、十分吵闹以及迅速变化的环境。
- 尽可能为各种工厂员工有效地服务，不论员工的性别、所说语言、语音和说话方式等等。
- “立即”对操作员的语音作出回应，以减少成本以及因推迟而给用户带来的麻烦。最小化“总使用成本”，包括要求的任何预先使用准备而损失的时间，以及为了用户在工作时因识别器推迟和错误而损失的时间。

语音辨识器如何工作

以电脑为基础的语音辨识器匹配模式。事件顺序（更加简化）如下：

- 识别器装载了一系列声音参考模式，这些模式代表着应用程式期待使用者说出的话语或者部分字（“音素”）。

技术专家采用术语“积极词汇”来表示期待用户在任何时候说出的话语清单，采用“词汇”来表示使用者在使用应用程式工作时可能说出的所有话语的清单。

- 随后应用程式向识别器传送一个未知的声音，该声音代表着用户说出的话语（话语或一系列话语），或者是外来的声音，也可能是受外来声音干扰的话语。
- 识别器将未知的声音“分类”，并汇报未知语音与一系列参考语音模式之间的最佳匹配。例如，如果参考模式代表的是数位元，那么识别器将报告，未知语音与数位顺序 123 最相配。识别器还将汇报未知语音与参考语音模式之间最接近的配对。若此类配对“得分”太低，那么应用程式将确定，未知语音最有可能属于外来音，而不是使用者的语音。在此情况下，应用程式将忽略报告的输出语音。

为何语音辨识如此困难？

若人们说话一致，那么语音辨识（通过人类和电脑）就是相对简单的问题了。但是我们说话方式并不一致。言语表达就像是雪花 - 没有哪两片完全一样。A 的说话方式很可能与 B 的说话方式差异很大。更糟糕的是，即便 A 连续多次重复单词“one（意思为一）”，每一次重复之间都有微妙的区别。语音辨识器还受到其他因素的进一步挑战：

- 当我们不停顿地说出多个单词时，我们说出每个单词的方式都受到其前后的单词影响。这就叫做协同发音。这也影响着单词的发音，因此可能采用不同的方式说出同一个语音，这取决于其前后的发音。
- 话语可能受到背景声音的干扰。
- 应用程式可能向识别器传送不包含任何使用者语音的外来音。
- 用户发出的声音可能无法准确地传送至识别器（例如，当两者之间存在电话连接时）。

当识别器出错时（人类和电脑都会出错），这些错误有三种—插入、删除和代替：

语音辨识错误示例		
错误类型	**说话者言语**	**别器识别的说话者言语**
插入	< 无 > 153	1 1593
删除	1 153	< 无 > 13
代替	153 5	193 9

在这三种类型的错误的情况下最大化识别器的性能实在是一个挑战，这点您可能也能预料到。例如，使识别器不那么容易受到插入错误的影响，就可能会增加删除错误影响的可能性。

简化问题

设计语音辨识器的人的重要目的在于，通过尽可能简化识别器的问题将识别器的错误最小化。有许多方法可以实现此目标，例如：

- 我们可以限制识别器的词汇。在听写系统中，我们可以运用的约束条件是十分有限的：用户在任何时刻几乎能说出任何话。在工业应用中，若系统要求使用者输入数量，我们可以通过告诉识别器来简化识别器问题，仅接受一连串的数字。
- 我们可以坚持在限制背景声音的环境下进行工作。这对于听写系统来说是合理的，但是对于旨在支援工厂或仓库环境下的员工的行业系统来说却是不切实际的。在这种环境下，我们可以做的是通过采用特殊的“消音”麦克风来最小化背景声音，并采用旨在最小化“噪音污染”影响的识别器演算法。

- 我们可以允许系统利用对使用者的了解。此外，不同的系统有着不同的应用理解的能力。
- 电话系统（例如，航班资讯供应商）实际上无法要求用户辨认，而且交易时间如此短暂，以至于系统在交易期间对使用者的说话方式知之甚少。
- 听写系统可要求新使用者在使用新系统之前，通过阅读长达五至十五分钟的一份或多份准备好的手稿来对其说话。这使得识别器可以获得关于用户“声音类型”（例如，声调高或声调低）和口音的信息。
- “较少词汇量”体系（例如，用于仓库中）可以要求新用户对其说出他或她在工作时可能说出的具体话语。随后系统就可以于特定使用者对应词汇中设定词的“声音范本”。这种利用对于使用者的了解的系统被称为“完全经培训”或者“与说话者有关”的系统。
- 一些识别器要求用户在每次说话（例如，“开始 1 2 3，停止”）前后都要说出“固定代码”。这些固定代码不仅可以改善识别器薄弱环节的性能，还增加了用户必须说出的话的数量，这将对工作效率产生负面的影响。在 Vocollect，我们已选择确保识别器提供最佳的表现，而不增添用户说出固定代码的额外负担。
- 最先进的识别器可以在他 / 她工作时对其进行了解。Vocollect 指的是“适应性识别”的技术，2006 年我们将该技术并入我们的产品中。

仓库语音辨识器的多种选择

当设计供仓库使用的语音辨识器时，有些决策很容易作出，但是有些决策却需要仔细考量。很明显，设计者应该：

- 限制词汇，以便与任务进行配对
- 采用设备（例如，麦克风）和计算方法来最小化背景声音的影响
- 避免使用固定代码
- 适应用户

剩下的主要设计决策就在于，使用未经培训（“与说话者无关”）或经培训（“与说话者有关”）的识别器。让我们对影响技术选择的仓库应用特征进行审查：

- 较少的固定词汇：未出现在许多其他语音应用中的这种特征允许采用完全经培训的识别器。
- 每位用户的多次汇报以及高汇报率：如以下所示，这些特征使识别准确率和反应速度变得十分重要，因为错误和推迟会迅速增加。
- 多语言、非本地工作人员：要求涵盖了广泛的语言、发音和说话方式。
- 必须适用于每位用户：没有切实可行的替代输入资料方案。
- 在嘈杂环境下说出的短语或者短字：短语或短字可导致插入错误，因此背景声音的不可透性是十分重要的。
- 不断变化的说话方式和背景声音：使用者说话方式的变化存在多种原因，例如，当用户下班后感到疲惫的情况下，说话方式就会发生变化。

下表回顾了仓库中语音辨识器的主要设计目标，表明未经培训或者经培训的识别器是否具有其固有的优势。

目标	经培训 ‘与说者有关’	未经培训 ‘与说者无关’
最小化提前使用的培训时间		✔
最大化准确率和工作人员的工作效率	✔	
适用于各种语言	✔	
不受口音、声音类型和性别等的影响	✔	
最大化地消除背景声音	✔	
最大化用户满意度	✔	
最大化适应性识别的好处	✔	

成本影响计算

显然，对使用何种类型的识别器，任何计算中都应考虑培训的成本以及改善业绩的好处。尽管我们不能轻易衡量出用户满意的好处或者成本，但是我们可以轻易地评估预先使用培训以及使用过程中出现错误的成本。

我们采用以下假设：

操作员的工资（薪水和福利）成本：	20 美元 / 小时
声音装置的使用：	一天 8 小时，一年 360 天
执行频率（例如，每个小时的拣选数）	200
每次执行所说的单词：	4
识别错误的工时成本：	3.5 秒
预先使用的培训时间：	20 分钟

关于以上假设的注释：

- 若仓库一周运行 5 天而非 7 天的话，声音装置的使用时间可能会更低，但是在多班制运行中使用时间会相对提高。
- 执行频率对应于典型的“批量拣选”操作。其他任务的执行频率可能提高，也可能降低。
- 四个单词通常是任何仓库运营中最少的：一个简单的“无一例外”的拣选执行中，操作员要说出三位元校验码，以确定拣选位置，说出一个数字来确认拣选的数量。
- Vocollect 已通过观察来衡量一个识别错误的工时成本。从错误中恢复的时间可能要大于此处所使用的资料，但是有经验的操作员有时候可以一边工作，一边从错误中恢复过来。

预先使用的培训时间资料代表着 Vocollect 仓库中的语音系统。考虑到以上假设，我们可以计算得出：

每天所说的单词 = 200 * 8 * 4 = 6,400

Vocollect 已花费大量的时间和精力，数年来致力于根据在仓库中采用我们的语音系统的使用者的说话实例创造一个资料库。我们利用该资料库，以便尽可能最好地评估用户在“真实世界”中可能看到的错误率。无论何时，当我们改善识别计算方法时，我们总是首先对资料库的改变进行测试，然后确认实地考察的业绩。专为仓库应用记录的资料集的使用为我们提供了实验室改进与实地考察汇报结果之间的良好关联。

从 Vocollect 的经验可以明确验证、但是却很难衡量的是识别器性能对员工满意度、整体员工的业绩以及设备滥用的影响。我们相信，此类“软”问题的存在不但真实，而且十分重要，这促使我们在改善识别器性能的同时，将注意力集中在除识别器计算方法以外的产品和服务属性，包括各种耳机设计和用户培训的问题。

测试结果

Vocollect 最近采用我们仓库内的资料库进行了测试，以便将我们培训的语音辨识器的性能与其他可用的多个未经培训的识别器（包括其他供应商仓库语音系统系统中最常用的识别器）的性能进行对比。

我们的测试结果表明，对于仓库使用而言，从经培训的识别器转移至未经培训的识别器而增加的单词错误率为几个百分点甚至更多。事实上，对于中等至重口音的说话人而言，增加的错误率从 6% 至 20% 以上不等。如以下图表所示，考虑到以上假设，使用的每个语音装置每年导致的成本增加可轻易达到 1000 美元（即便是在单班制运行中），而在执行多班制的工厂中每年的成本可能达到几千美元。

值得注意的是，本档中的分析仅适用于仓库中的语音系统。例如，Vocollect 已开发并采用未经培训的识别器，以供我们的医疗业务之用，在此情况下应用程式具有十分不同的属性。但是我们还是相信，经培训的识别器依旧是仓库使用中选择的产品。

因此，“错误率”每增加 1%（例如，识别器出错的次数从 1% 上升至 2%），我们可以得出：

每天增加的错误 = 6,400 * 1% = 64

每天损失的时间 = 64 * 3.5 = 224 秒 = 3.7 分钟

每年损失的时间 = 3.7 * 360 / 60 = 22.4 小时

每年的成本 = 22.4 * 20 = 450 美元

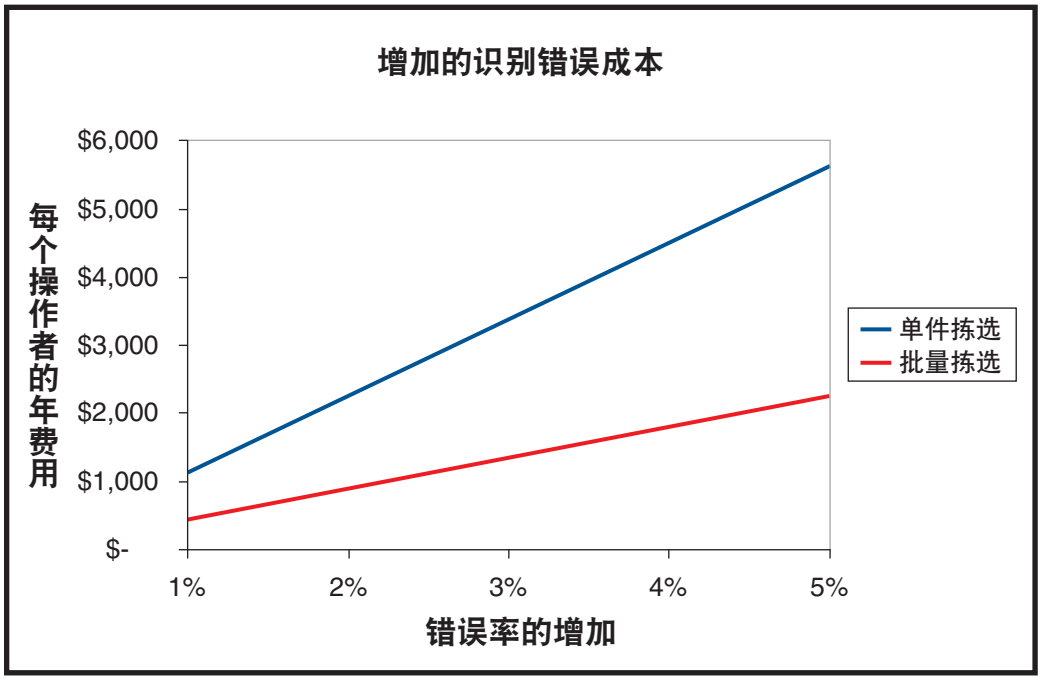
注意，此处所说的错误可以是前文描述的三种错误类型中的任何一种。未经培训的识别器尤其容易受到插入错误的影响。

因此，若采用经培训的识别器，也只能将单词错误率降低 1%。在预先使用的培训期间投资的回报期不到 6 个工作日。每个操作员每年节省的成本为 450 美元。

从上文中我们将看出，1% 只是对经培训和未经培训的识别器之间的差异的十分保守估计。

衡量业绩

要采用有效的方式来衡量识别器的错误率相当困难。公司可能声称其识别器准确率达 99.7%，而且也很可能设计出一个测试，测试结果显示出几乎所有的识别器都达到了 99.7% 的准确率（错误率为 0.3%）。但是也很可能为同一个识别器设计出一个不同的测试，显示出 10% 或者更高的错误率(30 x 更差的性能)，即便是简单的“是”或者“否”的任务。因此，识别器的准确率声明应考虑到过高的预估。公司的最佳做法是，首先采用大量资料集以及其用户经验的结果进行衡量。这个过程代价高昂且耗时。随后，有必要创造一个测试环境，以尽可能接近真实结果的方式进行复制。现在，若一个人对识别器进行更改并发现测试结果有所改善，那么实际用户也将实现类似的改善。有必要定期通过收集更多真实资料，采用最先进的识别器，并将结果与测试环境下的结果进行比较，从而重新进行检验。即便是如此严格地执行，也很难做出有用且可靠的数量准确率声明。例如，在消除背景声音方面的重大改进对于不含此类背景声音的环境丝毫没有影响。



注意：假定单件拣选率为每小时 500 条生产线，批量拣选率为每小时 200 条生产线。

结论

在仓库应用中，未经培训的识别器具有的好处在于：用户无需在培训上投入时间。但是从长期来看，经培训的识别器将产生更多回报。应用程式的特征不仅允许在仓库中采用完全经培训的识别器，同时也是设计识别器（尤其是为仓库而设计）的人员的最佳选择。首先，他们提供了更高的准确率，因为他们可以更好地区分和识别个人说出每个单词的方式：他们无需考虑一个区域或者某种语言的发音变化。这种专门化的特征可更容易地使其拒绝不应识别的语音，防止产生代价高昂的插入错误。

此外，加班时间内产生说话方式的变化使适应性识别成为仓库应用的选择。尽管完全经培训和未经培训的识别器都可能具有适应性，但是采用完整单词模式的完全经培训的识别器，相对于未经培训的识别器来说，通过采用适应程式可实现更高的准确率，其中未经培训的识别器采用的是以音素为依据的模型（单词中的独立发音）。

来自许多国家、说着各种语言、多种方言和口音的 30 多万使用者对 Vocollect 语音系统的广泛认可，已证明了 Vocollect 经培训的识别器方法的成功。

总之：

1. 仓库中应用语音技术的属性是支援采用完全经培训的识别器，而不是未经培训的识别器。
2. 在仓库应用中，培训识别器的最低成本相对于培训带来的业绩改善来说是更具有价值的。
3. 相对于未经培训的识别器来说，经培训的识别器每年为每位操作员节约的操作成本从几百美元到几千美元不等，这取决于应用程式的属性以及相关的识别器性能。
4. 经培训的识别器在支援多语言工作人员方面提供了大量额外好处。

关于 Vocollect

Vocollect, Inc. 公司是语音解决方案提供商的全球领导者，致力于为配送和仓库环境中的移动工人设计和提供语音导向解决方案。凭借语音识别软件和高度词汇差异语音识别准确性，客户可以通过语音实现更高水平的绩效。每天，全球有超过 30 万名工人使用 Vocollect 语音解决方案，将价值超过 30 亿美元的商品从配送中心和仓库配送到客户所在地。 Vocollect Voice 产品由一支汇聚了 2,000 多名供应链分销商和渠道合作伙伴专家组成的全球团队提供支持。 Vocollect VoiceWorld 套件可与各大 WMS 和 ERP 系统（包括 SAP AG 解决方案） 集成，并支持业界领先的移动计算终端。

欲知详情，请访问：www.vocollect.com

Vocollect APAC:
apac@vocollect.com

Vocollect® a business unit of Intermec®

Published by Vocollect, Inc.
703 Rodi Road, Pittsburgh, PA 15235
(412) 829-8145, Fax (412) 829-0972, http://www.vocollect.com