〔19〕中华人民共和国专利局

〔51〕Int.Cl.⁴
G06F 15/38

〔12〕发明专利申请公开说明书

〔11〕CN 87 1 06964 A

CN 87 1 06964 A

〔43〕公开日 1988年6月1日

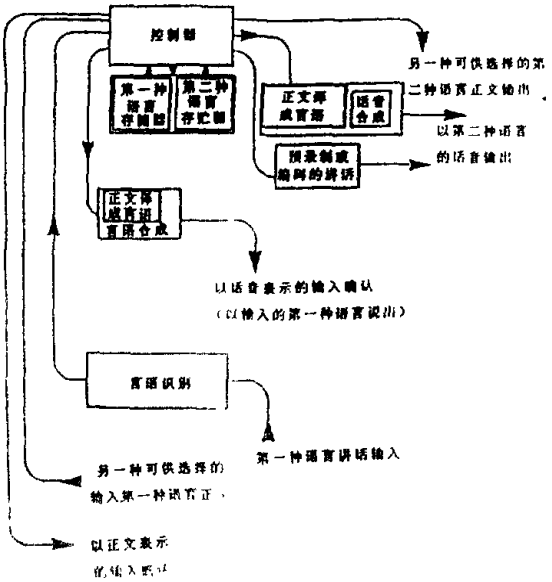
[21]申请号 87 1 06964
[22]申请日 87.10.3
[30]优先权
[32]86.10.3 [33]GB [31]8623839
[32]87. 5.1 [33]GB [31]8710376
[71]申请人 英国电信公司
地址 英国英格兰伦敦
[72]发明人 弗雷德里克·沃里克·迈克尔·斯坦
蒂福德 马丁·乔治·斯蒂尔

[74]专利代理机构 中国专利代理有限公司
代理人 许新根 曹济洪

[54]发明名称 语言翻译系统

[57]摘要

一种将词组从第一种语言翻译成第二种语言的翻译系统包括一存储一批第二种语言词组的存储器。以第一种语言输入的词组个个都根据一个或多个关键词加以鉴定，然后以第二种语言输出相应的词组。这种短语集法使翻译能有效地快速而准确地进行，口译也不例外。因第二种语言的词组是事先编制好且存储在存储器中，故不会出现翻译质量差或不合语法结构等问题。输出可以是正文形式或采用言语合成以话音形式输出。



权　利　要　求　书

1．一种将词组从第一种语言翻译成第二种语言的系统，其特征在于，该翻译系统包括一存储一批第二种语言词组的存储器、一接收第一种语言词组的输入装置、一以第二种语言输出包括所述一批词组中一词组的词组的输出装置、一确定所述一批词组中哪一个词组对应于所述输入词组的鉴定装置和一响应所述鉴定装置来控制输出装置并确保从所述一批词组中输出对应于所述输入词组的装置。

2．如权利要求1 所述的系统，其特征在于，该系统包括一通过一数据线路可操作连接的第一和第二终端，第一终端包括所述输入装置和所述鉴定装置，第二终端包括所述存储器和所述输出装置，其中所述第一终端还包括一产生信息并通过所述数据线路传输到所述第二终端的信息发生装置，该信息则指示所述一批词组中哪一个词组对应于所述输入词组。

3．如权利要求1 或2 所述的系统，其特征在于，所述鉴定装置根据输入词组中存在的一个或多个关键词或关键词部分确定各词组的对应关系，所述鉴定装置应用第一个词汇表确定输入词组对应于所述一批词组中的哪一个词组或词组群（如有的话），当发现输入词组对应于未确定的其中一个词组群时，所述鉴定装置就应用第二个词汇表来确定输入词组对应于哪一个词组群。

4．如以上任一权利要求所述的系统，其特征在于，所述鉴定装置包括一言语识别器。

5．如以上任一权利要求所述的系统，其特征在于，所述输入装置能接收口头讲话输入，且所述输出装置提供话音输出。

6．如以上任一权利要求所述的系统，其特征在于，设置有一能将所述输入词组的某些部分不经翻译传送到所述输出装置，作为第二种语言词组的一部分输出。

7．一种实现以上任一权利要求所述的系统的设备。

8．如权利要求7 所述的设备，其特征在于，该设备还包括一将输入信息提供给所述输入装置的键盘和一以第二种语言提供正文输出的装置。

9．如权利要求7 或8 所述将第一种语言翻译成多种第二种语言中的任一种语言的设备，其特征在于，在一存储器中配备有所述多种第二种语言中每一种语言的另外一批词组。

说　　明　　书

语言翻译系统

本发明涉及一种将词组从第一种语言翻译成第二种语言的系统，特别是（但不是唯独地）涉及一种根据第一种语言的讲话产生第二种语言的话音的系统。

对快速自动语言翻译机特别是口译机的研究已有多年的历史了。但尽管近年来电子计算、言语识别和言语合成等技术取得了巨大进展，这类机器仍然只是梦幻和小说的资料。

迄今，对自动翻译正文(text)的计算机系统是进行过大量的研究工作的。但除了在为数不多极其有限的应用领域（例如天气预报的翻译）外，都未能取得能自动精确地翻译因而可以取代翻译员的这样的成果。言语识别的误差、再加上语调、重音等方面的信息和言语本身的不准确，使言语的翻译问题变得复杂化了。

现有正文语言翻译程序包又偏偏都是不完全的，不是在这方面欠缺，就是在那方面不足，不能满足言语至言语的翻译系统的要求。迄今，这类程序包多数是作为职业翻译员的辅助工具而设计的，所产生的输出必须经过其后编辑处理才能以它们的目标语言表示出来。多数程序包不是按项目单加以驱动和进行人-机对话，就是以缓慢的成批处理方式工作，两者都不适于“实时”言语操作。此外，翻译程序包也往往不可靠，原因在于，习惯语和其它例外情况最容易得出错误的输出：使用者得不到输出能正确予以翻译的保证。再者，现有系统中用的中央处理机非常密集，这使它们的运行费用非常贵，因而不适用于许多对费用极其敏感的用途。

本发明试图提供一种上述不足和缺点有所减少的翻译系统。

根据本发明提供了一种将词组从第一种语言翻译成第二种语言的系统，该系统包括：一存储一批第二种语言词组的存储器；一接收第一种语言词组的输入装置；一以第二种语言输出包括所述一批词组中的一词组的词组的输出装置；一确定所述一批词组中哪一个词组对应于所述输入词组的鉴定装置；一响应所述鉴定装置来控制输出装置并确保从所述一批词组中输出对应于所述输入词组的装置。

这种系统翻译速度非常快，所需要的时间主要花在识别／鉴定输入词组和查找第二种语言中的“答案”。

该系统还可设计得可以给使用者提供输入确认信号，表明系统已正确识别／理解使用者的意思，这在言语翻译系统中当然是特别重要的。

由于所存储的一批词组是仅仅由预先做好的准确翻译组成的，因此一旦使用者获得他的信息已被正确鉴定的确认信号时，翻译的准确性即得到了保证。

这种翻译系统还可能同时将一种语言迅速地翻译成若干种第二种语言，所需要增加的部件基本上只是另外一些存储另外每种第二种语言的各批词组的存储器。

现在参照附图介绍本发明的一些实施例。附图中，图1 是显示根据本发明的系统各主要组件的方框图。

本发明是以我们的下列见解为基础的，即用数目小得多的关键词有可能鉴定和抓住大量截然不同的词组的语义内容。通过适当地选择关键词，就有可能使用现行市面上出售的只能识别数目远比大量有用的词组组中所包含的词少得多的词的言语识别器，来鉴定和区别大量的词组组。

因此翻译系统的性能总的说来有赖于那些关键词正确辨别词组的能力。识别词组的能力越大，系统在识别上的容许误差就越大，同时说话者本人所造成的容许误差也大。

关键词的选择

合适的检索程序如下：

1．根据各词在词组中出现的频率整理N个有关词组中K个词中的每一个词。

2．选择M个最频繁出现的词列入基本关键词表中，其中M为言语识别器词汇中的词数。

3．然后确定各词是否存在于各词组中。清点计数不为关键词所鉴别的词组的数目(E)。

4．令i-1。

5．临时从列表中删除一关键词，计算E的新值(E')。

6．把记分E'-E指定给被暂时删除的关键词。这是在关键词更新之后对系统性能变坏的程度的一个衡量，也就是它对系统总性能的影响程度的一个衡量。〔实际上，这种衡量是用以确保各关键词尽可能多地识别词组对而不致仅仅重复其它词组对的功能〕。

7．将暂时删除了的关键词复位，对M个关键词的每一个关键词重复上述过程。

8．从现有关键词表中除去记分最低的词。

9．用第M+i 个最频繁出现的关键词代替该除去的词，然后计算新的E值。

10．若新E值所表示的性能比原E值的好，就增加i 值，同时除非M+i>K，过程停止，否则从第5 步起重复过程。要不然，将第M+i 个词放弃不用，增加i 值，同时除非M+i>K,在此情况下，在第8 步最后一个被除去的词复位，过程停止，否则过程就从第9 步起重复进行。

最后得出的关键词表包含识别词组用的由M个关键单词组成的最佳关键词组。

从上一次重复的最佳M个词开始的进一步重复可以进一步改进词组

的识别效果。第1 步中整理待选用词的系列可采用试探法而不用频率法，特别是有既定的语言信息可资利用时更是如此。此外，可能在出现频率表中趋近底部的词对词组识别的帮助不大，因此是不值得超出出现频率表中前三分之一或前一半的范围进行检索的。

往往有这样的情况：在检索过程中，大部分词组都辨别出来了，而E过早趋近于零。在这些情况下，通过根据只有当不同的关键词多于一个时才可认为该诸词组已鉴别出来了来计算E的情况，可以获得进一步的改善。这样可以确保用多于最起码数目的关键词来识别大多数词组，同时在某种程度上起防止言语识别误差的作用。

在检索过程中会愈来愈明显：有些类型的词组如果不扩大关键词汇是永远也识别不出的。这些词组“簇”或词组群的区别往往只在于其中一个字或一连串辅助字（例如商业信件中的日期）的不同而已，而且都是自动得出的选择对象，用以编制关键词辅助词汇（下面将详细介绍）。

显然识别单个关键词时并没有考虑词序和可能包含的附加意义。因此，各词之间具有各种识别可能性的关键词对（或其它复合词）的有无也可用以改进单个关键词组的有效性。这给言语识别带来了这样的好处，即可以提高识别性能而无需增加识别词汇。应用在正文上时，还可以通过对关键词的归纳推广使其包括标点符号、词的各部分和词与词各部分的组合等而获得进一步的改进。例如，“-ing*bed”（其中* 可以是任何词）可出现在“making the bed”和“selling a bed ”中。

若能进一步解决词组的混淆问题，则采用关键词对（例如We**to）可以提高作为组成部分的单词的价值。检索一些不一定邻接而为不同数量的其它词所隔开的词对也得从按出现频率进行整理的准备工作开始。其组成词都是在M个关键词中的词对，若它们能解决任何遗留的词组的混淆现象，则可以整顿好的列表中制取。和先前一样，对最后得出的单个关键词和关键词对列表进行记分，并计算总的词组混淆记分值E。

现在开始对其中一个或两个组成关键词都不在现有关键词表中、使用情况更好的词对进行检索。从按出现频率整顿的序列顶部选取下一个词对候选对象，然后添加到关键词表中。同时也加入所添加词对中还未出现的单个关键词，并删除等量使用情况最差的单个关键词。这一下，其它词对，若它们的组成词不再存在的话也可能被删除。计算 E的新值(E')。若情况有所改进且E'<E，就保留对关键词表所作的最新修改，否则将列表回复到其以前的状态。其它词对就根据出现频率的次序进行处理，尽管和检索单个关键词时一样，但可以采用其它试探法提供词对候选对象。

值得注意的是，有些关键词通过加入到若干词组而比它们单独使用时对总性能的影响要大。

该方法可推广应用到更大的关键词的分组(>2 个词) 上，但由于出现频率的减少，所以只有在词组是极其大量时，解决词组混淆问题的作用才突出。

检索关键词的计算量随着关键词和词组数量而增加。这可通过先对混淆的或极其接近混淆的词组子集合进行运算来减少。如此得出的关键词及其记分，对将以更完整的词组组进行工作的主要算法，提供了更有效的候选关键词的次序安排。

在言语识别应用中，一些不在关键词组中的词会产生许多虚假的关键词识别，例如出现“ I”一词时可能总是会被识别为“by”。但若在开始检索关键词之前和在接着的词组识别时把混淆词群视为同义词性质的，则真正识别词组时就不应受此问题的影响。此外，由于这类同义词总的出现的频繁性必然高于个别词的出现频率，因此通常有较大的词组信息量与它们的检测过程有关的。

关键词的应用可以推广到重复出现频率更高，且所带的词组辨别信息比整个字更多的关键词各部分(例如音[illegible]上。此外，在连续言语中

识别某些词的部分是要比识别完整的词更容易，因而在接收连续言语话音输入的翻译系统中更受欢迎。在本说明书中，“关键词”一词是整个关键词和关键词的部分的简称。

有许多种词组彼此间的区别仅在于其可能包含日期、时间、价格、项目、名称或其它词群等细节的附属词组和附属子句的不同。有可能一个言语识别器的词汇足以确定一个词组为一特定的词组类别或词组群，但其容量还不足以容纳足量的关键词以识别各附属结构。此外，很有可能识别词组类别和附属结构所需的全部词汇包含更多易于混淆的词。这就是说，即使识别器的容量足以容纳整个词汇，但由于性能低，因而不能可靠地识别词组和附属词组。本发明的方法具有这样的优点，即可以将原始的发音或原始发音的某些变音存储在缓冲存储器中，而且一旦词组类别被确定下来后，可以用预期在对该词组类别来说系特殊的附属词串中有的关键词组来重复识别过程。这样，识别器就完全无需马上应付具有许多潜在的词混淆情况的整个词汇，但在使用者看来似乎如此。应该指出，辅助识别过程的速度并不受原始发音速率的限制，且一般可进行得比实时快很多，因而不需要引入显著的延迟。识别过程可按需要重复多次，以识别所要求的词组及其基本结构。这样，由于词组是在多个识别阶段中鉴别的，且识别器在各阶段对关键词的不同词汇进行提取，因而有可能“套用”识别过程。

附属词串有很多（虽然不是全部）在源语言中与上下文无关。这是因为只有当若干替换词可能建立密切的上下文依赖关系，且任何附属词串这样做的可能性更小时，才能确定附属词串的位置。此外，上下文关系的重要性意味着在可能成为附属词串内外的词之间存在着依赖关系，因而无需采用词串内的词辨别整个词组的关键词是有一定的范围的。例如在包含变更日期的词组中（在这种情况下，除日期本身之外，词组中是很少需要修改任何词的）就是如此。（未来的研究课题应该论证“这

种与上下文无关的性质在各语言之间是不变的”这个假设 并用它来无限制地扩大短语集的翻译。）

本发明的这种特点当应用到正文翻译时有显著的好外：可以采用较小词典或短语集的类似体系来大大削减查大词典所需的计算费用。某些附属词组是无需翻译的，而且往往在这些情况下通常不可能自动识别这些词组中的词。这种最普遍的情况发生在参照象专有名词之类的称号发音时发生的，例如：“我可与史密斯先生说话吗？”。和前面所述的一样，本发明的系统能识别词组类连同各词在缓冲存储器中对应于称号标记的位置。于是在翻译过程中对这类称号标记字的处理仅仅需要在适当位置将原始话音信号以目标语言发音发送即可。显然最好是合成出来的目标语言的声与原说话人的话音一致，而且正文转换成言语的合成器的要求是，某些言语话音参量可以调节，使得尽量满足话音的一致性（例如老年人/年轻人，男人/女人）。

为使使用者有把握肯定译出来的目标语言词组的正确性，翻译系统指示出它将翻译的输入语言的词组。为做到这一点，翻译系统配备有存储输入语言词组的所有组成部分的存储器。

各词组在系统中最好作为正文以例如美国信息交换标准码(ASCII)编码的形式进行存储，因为这样可以比普通压缩扩展式或非压缩扩展式言语大大减少所需用的存储器数。需要输出讲话时，可以从存储器中取出正文，把它传送到正文/言语转换器和言语合成器中。采用ASCII 编码正文存储器时，1 字符需要1 字节，这就是说，用半兆字节的存储量即可存储10,000个词组左右。因此能翻译10,000个词组左右的系统约需要1 兆字节的存储量，这些是不难存储在硬磁盘上的。双向通信有可能采用两个结构对称的翻译系统。这样做的好处是，各系统只需要识别和合成以操作该单元的人员的语言表示的词。用第二单元的通信要借助一个规定词组和任何附属词组内容的协议。该协议与语言无关，因而可以

无需识别目标语言来传输信息。此外，这样做还可采用多种不同语言同时接收从单个单元输出的译文。

实例

操作一种接到电话网络的示范系统，用以示范查找短语集方法的可行性。该示范装置采用一Votan 言语识别器、一Infovox 言语合成器和一IBM 公司的XT型个人电子计算机。

Votan 言语识别器能识别多达64个在电话网络上连续说出的词。考虑到象“yes ”，“no”，“quit”和“enter ”之类的系统控制词，可选多达60个词作为关键词。输入词组中不容许有任何一个系统控制词出现，所以可能的话最好采用控制按钮或控制键而不用口头指令。

存储的词组包括400 个英语词组及其法语的等效词组。

英语词组约含1100个不同的词。这些数目相当于普通商业用语短语集一般所包含的词组数目的范围。

按上述原理操作关键词抽出软件之后，选用了60个关键词，该60个关键词成功地识别了所有词组。在400 个词组中，只有32个词组只用单个词辨别（该32个词组为16对词组）。

示范系统一经识别关键词就访问有关词组，（口头）向使用者确认，然后通过正文/言语合成器输出等效的法语词组。

应该指出的是，正文/言语合成器并非本发明的主题。目标语言可以从预录制或编码的词和词组合成，这样做是完全行得通的，也是确实有益的。这样做的好处是，这类言语可由使用者录制，因而在话音方面会与任何记入的话音一致，而且可以无需将正文合成为言语语音。这种方法还可以在一些重要国家的语言中无需将正文合成为言语话音，在这些重要国家中这种技术是不可能在最近的将来制造出可使用的硬件的——例如印地语和阿拉伯语。

除将言语翻译成另一种语言的言语外，本发明当然也适用于把正文

翻译成另一种语言的、正文、将正文翻译转换成另一种语言的讲活或将讲话翻译转换成另一种语言的正文等的翻译。特别有用的用途是在办公室自动化的领域中，在该领域中是不难实现将讲话直接翻译成外语正文的翻译机的。实质上，这种翻译机可以采用言语识别器、上述软件和控制系统，但将第二种语言正文输出到一打印机上或电传机或其它电信线路上。当然用几种语言编制日用商业通信的普通词组并不是什么难事。

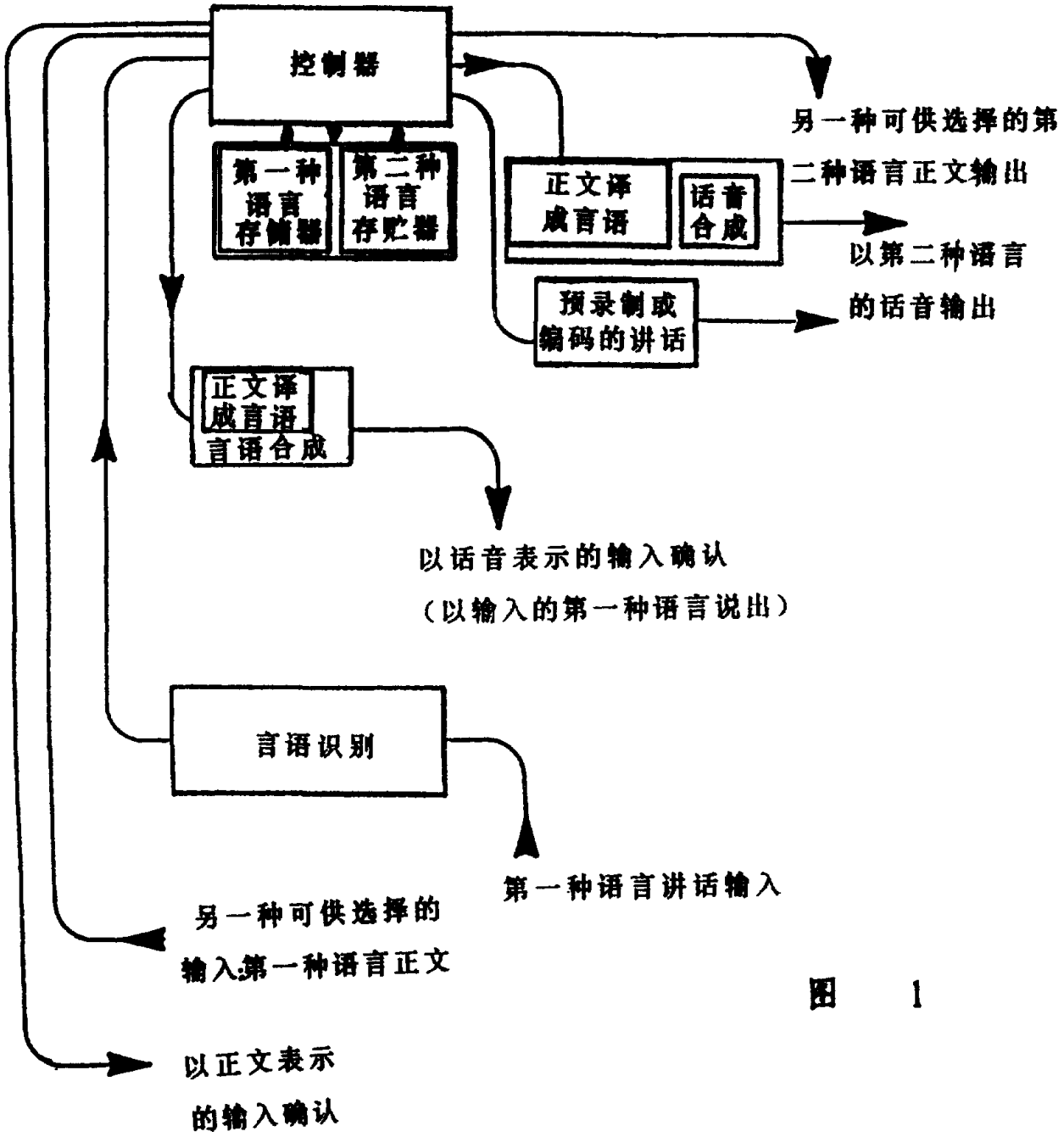

控制器
第一种语言存储器
第二种语言存贮器
正文译成言语
话音合成
另一种可供选择的第二种语言正文输出
以第二种语言的话音输出
预录制或编码的讲话
正文译成言语言语合成
以话音表示的输入确认
（以输入的第一种语言说出）
言语识别
第一种语言讲话输入
另一种可供选择的输入第一种语言正文
以正文表示的输入确认

图　1