汉语口语语料库研究的若干问题

王显芳　杜利民　（中科院声学所语音交互信息技术研究中心）

[摘要]　结合国际上口语语料研究的最新进展和作者的研究经验，介绍并讨论有关汉语口语语料库研究的若干问题。

关键词：　口语语音识别　口语语料库

1　前言

设计口语语音识别系统就必然要收集口语语料。随着口语人机交互系统技术的不断发展，语音数据的存储和管理在系统的设计中变得越来越重要[1]。口语语料库数据量大，语料的收集、整理、标注和维护工作量也很大，因此需要一个数据结构合理、使用方便的数据库管理工具来存储和管理口语语料。关系数据库管理系统（RDBMS）被认为是一种理想的语音语料库管理工具[2]。语音和文字之间的对应关系以及其它相关信息能够被充分地表示出来，也能很方便有效地存储和查询。通过建立其它的表，关系数据库还可以表述说话人、语言、语音特征等各个层面的内容。

建立一个口语语料库的目的是为了更好地进行口语语音识别，因此，口语语料库中的数据应该能够充分反映口语现象。口语是人们在进行自然交流时的语言现象，而在实际语料收集时，人们总是自觉不自觉地改变了讲话的习惯，使得收集到的语料不再能够充分反映口语现象。如何才能收集到既能反映人们口语现象又能限制在一定任务领域内的实际语料是个非常值得探讨的问题。

实际口语语料是为了提高计算机的语音识别正确率。因此还应该充分研究人和计算机讲话时的语言现象。由于心理预期的作用，人们在同计算机讲话时的语言习惯与人人交互的语言习惯往往是不一样的。因此在收集口语语料时，不仅要收集人人对话时的语料，而且还要收集人机对话的口语语料。但要收集人机对话的口语语料必须要在一个实际可用的口语对话系统开发出来以后才能进行。文献[8]中提到了利用WOZ（Wizard of OZ）方法模拟收集人机对话语料的方法。在口语识别系统开发和研制的前期阶段可以采用这种方法。在实际可用的原形系统开发出来以后，再大量收集实际的人机对话语料来改进系统性能，提高识别率，并从收集到的语料中总结出人机交互时的语法现象和语言习惯，使得人们同计算机交互更加自然、方便，人机界面更加友好。

在收集到语料后，还必须进行标注等整理工作。首先要对口语中的非语言现象有一个明确的可操作的分类。这是一个很难确定的问题。因为人们很难说清口语中的非语言现象到底有多少种，而且怎样确定标准才能让进行实际标注工作的人员从始至终按照一个标准进行语料的标注也是一个非常难以做到的问题。此外，还要开发一个使用方便的集成维护工具以便能够快速有效地进行语料标注和数据管理。

中科院声学所语音交互技术研究中心在进行高鲁棒性的口语语音识别系统研究过程中进行了口语语料的收集工作，建立了汉语口语对话语料数据库系统CAS-IIS CSDDBV1.0，目前正在进行实际的语料收集工作，本文抛砖引玉地介绍有关值得注意的一些问题。

2　口语语料库的数据库结构

口语语料库的设计主要出于以下考虑：首先，语音文件和标注文件要分开，这样便于文件

的存放和拷贝，也便于收集到语音后统一标注。语音文件和标注文件的对应关系在关系数据库中定义，在进行系统训练和测试时从数据中读出；其次，每一条语音文件都有一个全局唯一的 ID，用以标识此条语音，在数据库其余部分均以此 ID 代替该条语音，这样在以后记录口语对话信息时就会非常方便；再者，除去语音信息之外，语料库还记录说话人的部分信息，比如性别、年龄、籍贯，在可能的情况下还可以记录联系方法，以备后用；最后，在语料库中除了记录语音所对应的文字信息外，还应记录语音的一些特征信息，比如背景噪音强弱、录音质量、口音、方言种类、说话人情绪、句子是否符合语法，以便在对系统训练、分类评测时使用。

3 口语语料库的数据收集

口语语料库的数据收集大致可以分为三个问题：一是语料的收集场景，即语料收集场所的问题；二是语料的采集方式即采音方式是麦克风方式，还是电话方式，这两种采音方式是口语语音识别系统实用的两种方式；三是收集的语料是人人对话（H-H）还是人机对话（H-C）。下面就分别讲述这三个问题。

3.1 口语语料的收集场景 口语语料库语料的收集有两种方式：一是实验室条件下模拟场景收集口语语料，二是在实际场景下收集实际口语对话语料。相比较而言，第一种方法简单易行，但收集到的语料中的语言现象可能和实际有些区别，不一定能够总结出实际对话中的语言规律。因为对话的场景和说话的方式是人们想象出来的，和实际对话肯定存在着一定的区别。而且因为是在实验室条件下，很难模拟出实际场景下出现的环境噪音和背景噪声，而这些在诸如汽车环境下口语语音识别、电话口语语音拨号、商场信息服务台等实用语音识别系统下有着重要意义。第二种方法可以收集到实际的语料，比较真实地反映口语现象。但这种方式收集到的语料往往需要很多整理工作，因为人们在实际对话时往往有很多聊天的内容，而语料库要求语料能够限制在一定任务领域内。这种方法实现时需要较多的环节，需要耗费大量人力和物力。而且实际操作的经验表明，因为涉及到隐私权问题，人们对这种语料收集方式有着一定的抵触情绪。在处理不好时还可能出现一些法律纠纷。

目前，语料库数据收集的方法是两种方法都采用，两种方式同时进行，都收集了一定的语料。经验表明，第二种方式在实际操作时难度较大。

3.2 口语语料的采音方式 语料的采音方式可以分为两种：麦克风方式和电话方式。这两种方式采集到的语音质量是不一样的。一般来说，麦克风方式采集的语音质量更好，因为原始语音未经任何处理即被采入，频谱上没有什么变化。而电话采集由于线路噪声以及电话线路传输带宽的原因，波形上会发生一些变化，尤其是通过手机，由于传输时经过压缩，现象更明显一些。在麦克风方式下由于话筒和说话者的距离发生变化的可能性较大，音场会发生变化，声音随着说话人的移动有着强弱变化，且环境噪声有时也会很大，在电话方式下这些问题基本上不会存在。在实际的语音识别系统中要根据两种不同情况做不同的预处理。

在通过电话方式采集时，必须使用电话语音系统。系统必须对各种电话事件如呼叫接入、呼叫建立、连接断开作出响应，需要一个电话控制设备。目前系统使用的是 Computer-Phone 。

3.3 口语语料的收集对象

一般来讲，口语对话系统语料库应该大量收集人机对话，因为设计口语对话系统的目的就是让人能够方便地以口语方式和计算机进行交互，人的说话对象是计算机而不是人。而在人机对话的时候，由于心理预期的作用，讲话基本上会正式一些，也不会是类似于聊天式的谈话，主题比较明确。但一个问题就是大量收集人机对话的前提是首先要有一个成熟可用的人机对话系统。这样就有一个次序问题。其解决的方法有两种：一是模拟场景的方法[3]，即模拟人机对话的场景，让一个人充当计算机的角

色。首先设计计算机的对话策略，这种策略可以是系统主导的(System-Initiative)，也可以是混合主导的(Mixed-Initiative)，这依赖于想模拟的人机对话系统的对话策略。然后研究用户可能的输入模式，对于可以归为一类的输入模式规定一种应答模式。即首先要设计出人机对话系统对话管理模块的实现方法。然后让模拟者严格遵守这些规则，并且在对话过程中要假设一些语音识别模块识别结果出现错误的情况。这种方法称为WOZ。虽然这种方法在一定程度上可以模拟人机对话，但毕竟是在“演戏”，和实际情况还是有一定的区别。第二种方法就是采用一种循序渐进的方法，首先利用一般的大词汇量连续语音识别系统作为对话系统的语音识别部分，然后设计出对话管理模块以及响应生成模块。利用此系统作为对话系统的原型来收集实际语料，然后根据实际语料来设计和训练新的语音识别模块，并用其代替原来的识别模块，如此循环下去，直至设计出鲁棒性较强的语音识别模块和语言处理模块。这种方法的一个缺陷就是在收集语料的初期，系统会出现很多识别错误和处理错误，显得非常“愚蠢”，可能导致用户的不满，从而采用不合作的方式和系统对话，难以收集到有用的语料。

目前收集人机对话采用的方法是将以上两种方式结合，在开始阶段采用WOZ方法收集一些语料，用这些语料作为训练数据来设计一个口语语音识别系统，然后采用第二种方法来逐渐提高系统的性能，并且在中间阶段采用两种方式并行，同时收集语料。这是较为实际可行的一种方法。

实际中还需要收集人人对话，这不仅对研究人们日常口语中语言现象有着重要意义，而且在有的系统中处理对象可能就是人人对话(比方说在口语翻译系统中)。收集实际的人人对话语料可以采用模拟场景的方法，也可以采用实际场景下录音的方法。一般倾向于实际场景下录音的方法。录音的任务领域可以是打出租车、银行业务、寻呼台呼叫服务、电话接线员、火车票机票售票及信息查询、气象信息查询、旅行社信息查询等等。实际场景人人对话的收集需要涉及许多环节，实际操作的难度很大。目前这方面的工作虽然已经开始，但进展缓慢。

4　口语语料库的数据存储

为了方便地对语料库中的数据进行使用和管理，数据的存储结构是重要的。好的存储结构不仅方便口语对话系统的训练，而且便于将不同时期、不同地点、不同领域的语料结合起来，形成更大的语料库，并且便于语料库数据的拷贝和分布式存取。一般情况下，不能够假设进行语料标注的计算机和进行口语对话系统训练的计算机有相同的目录结构，也不能够假设它们具有相同的文件系统，而且在数据进行拷贝时，数据存储的目录结构一般都会发生变化。存储数据信息的文件和存储语音文件以及标注文件往往也存储在不同的计算机当中。采用存储文件绝对路径的方法缺乏扩展性和适应性[1]，所以在语料库中，不应该存储文件的绝对路径。

5　口语语料库的标注

语音语料收集之后，必须进行语料的标注，将语音数据和其文字内容对应起来。语料标注的目的是为了进行识别系统的训练，因此语料标注的方法和进行识别系统训练的方法是密切相关的。在识别系统训练时需要什么信息就需要在语料标注时予以标注出来。需要指出的有两点。(1)在语料标注时不加时戳信息，在进行口语识别系统训练时采用预处理算法将音和字对应起来。这种方法目前是比较普遍的，也能够很好地保证语料训练的正确性。这样在标注时就减少了很多Alignment的工作。(2)由于标注的语料是口语语料，口语语料中有很多的非语言现象(Non-Speech Event)，这些非语言现象在声音信号层面上和一般的语音难以区分，在标注的时候必须要标注出来。

非语言现象的识别在口语语音识别中是一个关键问题。目前连续语音识别系统用于口语语音识别时，大部分的识别错误都是由非语言

现象引起的。因此如何能够在识别时将非语言现象识别出来是个非常重要的问题。因此在语料标注时必须将非口语现象标注出来,以便在系统训练的时候能够总结出非语言现象的规律。目前采用的方法是对非语言现象进行分类,每一种非语言现象给出一种标注,标注的时候将其视为一个字标注出来。非语言现象大致可以分为以下几种:[UM]、[BREATH]、[SIL]、[LAUGH]、[NOISE]等。在英语等西方语言中还有词间中断又重新开始的现象,但因为汉语的最小单位是单音节的汉字,因此不存在这个现象。

非语言现象的分类也是个很难确定的问题。这和识别时将其进行划分的情况有很大关系。很难准确地说汉语口语中有多少种非语言现象。而且也难以保证每个人对于非语言现象能够采用同一个标准来划分,即使对于同一个人在不同的时间也可能有不同的标准。因此非语言现象的划分不能划分得太细,划得太细就会因为标注时标准的不统一而导致事实上的标注错误。但也不能划分得太粗,否则就失去了分类的意义。目前采用的方法是先准备一部分语料进行标注,在标注的时候发现难以确定的现象再进行讨论,直至可以确定一个保持不变的标准,在以后的语料标注中都采用同一个标准,不再更改。假如必须要改的话,就必须将标注过的语料重新标注一遍,以保证标注标准的唯一性和标注的正确性。同时在制定标准的时候还必须考虑到标注人员能否正确把握标注标准,并且要反复训练,以实现标注的正确性。

标注中还有一个难以解决的问题:在录音时有可能将对话双方的话同时都录下来,比如说在电话系统中将系统的响应和用户的言语同时录下来,这正是所谓语音识别中"Barge-In"技术需要解决的问题。因为不能将其中系统的响应划为非语言现象。但它又确实不是识别系统的合法输入。目前还没有找到好的解决办法。

6 数据管理工具

口语语料库还需要有一个使用方便的标注工具和维护工具。目前系统实现了一个集成的可视化图形界面标注工具,对于语料数据库有浏览、添加、删除、修改以及批处理添加等功能,尤其批处理添加功能能够将多条数据一次性加入数据库,避免了标注人员大量重复的劳动。系统还实现数据库维护界面和标注界面的紧密结合。可以在看到声音波形的情况下对语料进行标注,并可以局部选中,反复播放,有利于提高标注的正确性。

7 结束语

本文结合国际上口语语料库研究的最新进展和我们在汉语口语语料库研究中的经验,介绍并讨论了有关汉语口语语料库研究的若干问题。希望能对国内同行有所借鉴。

参 考 文 献

① Laws, M. and Kilgour, R. MOOSE: Management Of Otago Speech Environment. ICSLP 98, Sidney.

② Data, C. J. An Introduction to Database Systems. Volume I, 5th Edition. Addison-Wesley Publishing Co.

③ Woosung Kim and Myoung-Wan Koo. A Korean Speech Corpus For Train Ticket Reservation Aid System Based On Speech Recognition. Europe Speech 97, Greece.

注:王显芳,博士生。主要研究领域为语音识别,对话系统。

杜利民,同前。

·产品介绍·

与新型点火线圈匹配的 **IGBT** 新一代汽车已开始普遍使用新型点火线圈(coil-on-plug),为了与新线圈相匹配,国际整流器公司推出 14A IRGS14B40L IGBT 器件。新产品的电性能足以媲美甚至超过同类产品。

IRGS14B40L 是专为亚洲汽车制造业而设计的。它以国际整流器公司创新的箝位(clamped)IGBT 技术为基础,在每个火花塞的线圈上配装一个 IGBT,采用新技术的点火结构具有更高的成本效益。该产品可作为切换装置,准确控制初级点火线圈断开时的电压,从而更精确地控制火花塞次级电压。