网络环境下的语音识别方法*)

韩纪庆 张 磊 郑铁然
(哈尔滨工业大学计算机科学与工程系 哈尔滨150001)

摘 要 随着Internet技术的广泛使用,出现了通过Internet来传输语音的新的通信方式——VOIP技术;由此产生了网络环境下语音识别的新问题,这是一个富有挑战性的研究课题。本文将讨论这种网络环境下语音识别的方法和技术。

关键词 语音识别,Internet,特征

Speech Recognition Methods over Internet

HAN Ji-Qing ZHANG Lei ZHENG Tie-Ran
(Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001)

Abstract As the use of the Internet becomes more and more widespread, there is a new communication way, i.e. Voice over IP(VOIP). Consequently, recognition of VOIP is becoming a new and challenge research work. In this paper, we discuss the techniques of speech recognition over Internet.

Keywords Speech recognition, Internet, Features

1 引言

随着网络技术的日益成熟,通过Internet网络来传递声音的IP电话技术发展迅猛,已成为人们日常交流的重要手段之一。随着IP电话的发展,有关如何在IP电话中进行语音识别的研究引起了研究者们的注意,已成为本领域的热点问题之一。

IP电话由于其工作方式的特点,在传输中存在一些额外的信息损失,如网络中传输的语音都是使用各种声码器,考虑到带宽的限制,所传输的语音数据要进行压缩编码,这样在编解码过程中存在着信息的损失。同时,在网络传输过程中,语音信号经过编码压缩后打包在网络中传输,一般的传输协议中语音包是基于不可靠的RTP层传输的,这样会存在丢包的情况,因而会导致接收方获得的语音信号的音质受损。此外,数据包在传输过程中,由于网络的拥挤,还会存在包到达的延迟。这一切是传统语音识别方法中所没有涉及到的问题。

通常,包的延迟并不影响语音波形的变化,语音识别系统可以在允许的时间内等待延迟到达的数据包,然后再进行识别,这样不会对识别性能造成太大影响。因此,IP电话语音识别中主要考虑的是语音压缩和丢包造成的影响,以及如何克服这些影响的方法。通过模拟IP电话数据进行识别实验的研究表明[1],由于语音编码造成的性能下降在15%~30%,而对于丢包率小于5%的情况,其所造成的性能下降小于10%。对于ITU规定的几种标准编码方式,G.729D、G.723.1、G.729E和G.729编码而言,与通常的语音相比,G.729D编码和G.723.1编码方式会引起识别率的较大下降。相对而言,G.729E编码方式引起的误识率最小,而G.729编码方式引起的误识率介于中间。一般而言,低比特率的编码方式带来的编码损失较大,因而引起的误识率也就较大。当丢包率大于10%时,随着丢包率的增加,系统的识别性能明显下降。但在实际中,丢包率一般都小于5%,因此,由于语音编码所引起的语音识别性能的下降要大于丢包时的情况。

本文将讨论网络环境下由于语音编码和丢包所导致的语音识别性能下降的解决方法。

2 声码器损失的克服

一般来说,对网络上的语音识别,其后端的模型训练和模式匹配方法,同传统语音识别中的方法没有什么区别。两者不同的地方在于前端特征提取方法的不同。通常的语音识别系统,其特征参数是从采样、分帧后的语音波形数据中经过短时特征分析后获得的;而网络环境下的语音特征,需要从经过声码器编解码之后的压缩数据中获得。

我们知道声码器由编码器和解码器两部分组成,它们分别处于发送端和接收端。编码器主要是对连续模拟的语音信号进行压缩,以适合在有限带宽的条件下进行语音的传输。解码器在接收端将压缩的语音解码还原成语音信号用于播放。因此,对网络中的语音进行识别,一种最容易想到的方法是:先对压缩后语音信号进行解码,然后按传统的特征提取方法重新对语音信号进行加窗、计算静态特征和动态特征等。不同的语音声码器有不同的设计方法,它们在带宽、计算复杂性、质量等方面区别较大。基于国际语音编码标准G.723.1的声码器是比较常用的一种,它的性能较好,解码后的语音信号听觉效果良好。采用这种方法进行语音识别的原理如图1下部的虚线框所示[2]。对语音识别来说,它并不关心解码后是否能恢复为时域上的语音信号,更关心的是所获得的特征参数能否与语音识别模型的参数相匹配。一些研究表明:经过声码器后语音信号明显发生了畸变,从而导致了识别性能的下降。因而这种方法不是较好的选择。

另一种方法是在接收端直接从压缩后的语音中获得特征参数。比如对G.723.1编码器,由于它是基于码激励线性预测

*)本文受教育部跨世纪优秀人才培养计划基金资助。韩纪庆 教授,博士生导师;张 磊 博士生;郑铁然 博士生。

CELP 的方法，因此可以从接收端获得的量化的 LP 频谱中，进一步推导出所需要的用于语音识别的特征参数。其中所得到的频谱包络和从原始语音中获得的相同，唯一不同的是它是被量化表示的。但有研究表明，这种量化畸变不会对语音识别性能产生严重影响。图1上部的虚线框中给出了这种方法的原理。采用这样的方法，避免了第一种方法中先还原语音信号，再重新计算特征时产生的较大的误差；同时它还节省了解码还原语音信号所需的时间。

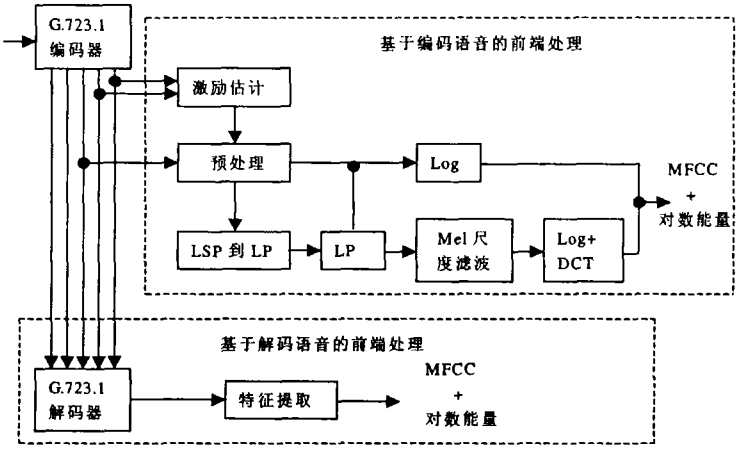

图1　两种基于 G.723.1声码器进行语音识别的前端处理过程

在网络中进行语音识别，特征除了可以采用 MFCC 等外，也可以根据具体声码器的编码方式，采用编码中利用到的中间特征，如 LSP 特征、LPC 特征或其它衍生出来的特征，如 PARCOR 系数、声道截面系数，以及倒谱系数等。

3　丢包损失的克服

对于丢包现象，在真实的网络环境中很难控制其丢失的多少。因此，为方便研究，通常是用一个模型来模拟可控制的丢包现象，如用 Gilbert 或 Elliott 模型来近似模拟[1,2]。图2给出了一个描述丢包现象的 Gilbert 模型，它是一个两状态的马尔可夫模型。这个模型包含两个状态，第一个状态是与低丢包率相关的 P_1状态，第二个状态 P_2对应的丢包率很高，即 $P_1<<P_2$。从第一个状态转移到第二个状态的转移概率用 P_t 表示，而从第二个状态到第一个状态的转移概率用 P_t 表示，则 $P_t<<1-P_t$。这样，从比较好的 P_1状态转移到状态 P_2的可能性较小。但一旦模型处于第二个状态，就不太容易转出第二个状态，这时就会产生大量的丢帧现象。

为处理丢包的问题，通常在识别前端加入对丢失帧进行检测和估计的方法[3-5]，其结构如图3所示。它包括两个阶段，第一个阶段利用包检测机制确定有没有丢包现象发生，如果存在丢包现象，则在第二阶段利用特征帧的一些特性来估计被丢失的语音帧。对丢失帧的检测主要是通过在特征矢量中加入一个反映帧序号的计数值，通过监测这个帧序号可以确定丢失帧的位置。此外，采用帧序号的方法还有利于对经过网络中传输后乱序的数据包重新进行排序。

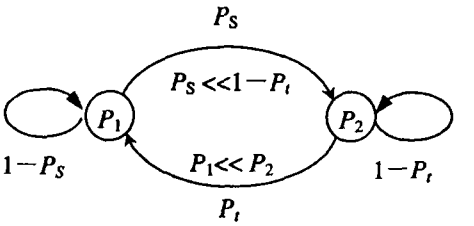

图2　模拟包丢失的模型

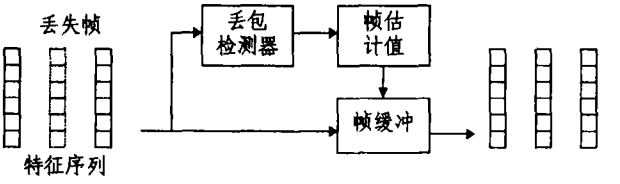

图3　网络语音识别处理中的前端处理阶段

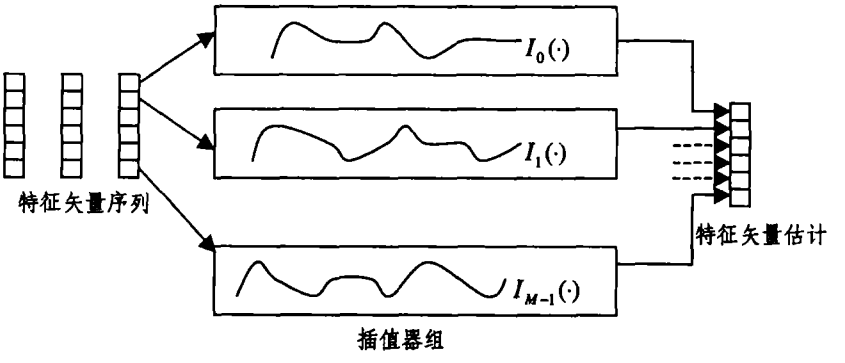

图4　丢失特征帧的插值

当丢包现象发生时，最简单的方法是用丢包前的一帧数据替代丢包帧的数据。较复杂的解决方法是应用一些插值算法，根据语音特征的轨迹来估计丢包语音帧的数据[4]。图4为

（下转第181页）

对30篇实验样本的综合评价表明，我们的方法在上述评价因子的得分上优于传统的无主题区域检测的文摘方法，且对于部分文风自由、主题灵活的文本，我们的方法要优于仅基于相邻段落语义相似性计算的方法。

结论及今后的工作 本文提出了一种基于主题区域发现的中文自动文摘的研究方法，该方法产生的文摘能在尽可能全面地覆盖原文多个不同主题的同时，显著地缩减自身的冗余，从而能有效地平衡两者之间的矛盾。在我们的实验评价过程中，此方法取得了比较好的评价结果。

当然，本方法也存在不足之处，如术语抽取的质量还有待进一步提高、聚类算法和聚类分析算法还需进一步完善等。对于这些不足之处，我们将在今后的工作中逐一改进。

参 考 文 献

1 王继成，武港山，等. 一种篇章结构指导的中文 Web 文档自动摘要方法. 计算机研究与发展，2003，40(3)：398～405

2 刘建舟，何婷婷，姬东鸿. 基于开放式语料的汉语术语的自动抽取. 见：第二十届东方语言计算机处理国际学术会议论文集，2003. 43～49

3 Nomoto T，Matsumoto Yuji. A New Approach to Unsupervised Text Summarization. In：Proc. of ACM SIGIR'01，2001. 26～34

4 Gong Yihong，Liu Xin. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In：Proc. of ACM SIGIR'01，2001. 19～25

5 Pantel P，Lin Dekang. Document Clustering with Committees. In：Proc. of ACM SIGIR'02，2002. 199～206

6 Mitra P，Murthy C A，Pal S K. Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions of Pattern Analysis and Machine Intelligence，2002. 1～13

7 MANI I. Summarization Evaluation：An Overview. In：Proc. of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization. Tokyo：National Institute of Informatics，2001

8 MANI I. Recent Developments in Text Summarization. In：Proc. of CIKM'01，2001：529～531

9 杨晓兰，钟义信. 基于文本理解的自动文摘系统研究与实现. 电子学报，1998，26(7)：155～158

10 Kaufmann L，Rousseeuw P J. Clustering by means of medoids. In Statistical Data Analysis based on the L1 Norm. In：Dodge Y，ed. Amsterdam，1987. 405～416

11 Rissanen J. Modeling by the shortest description. Automatica，1978(14)：465～471

（上接第176页）

通常使用的用于估计丢失帧的插值方法。将特征矢量序列 $\{X_0, X_1, \cdots, X_N\}$ 输入到插值器组，特征矢量的每一维单独使用一个插值器，如第 m 维使用 $I_0(m)$。这样根据特征矢量轨迹信息可以估计出丢失的矢量 X_n，其第 m 维的估计 $\hat{x}_n(m)$ 可估计如下：

$$\hat{x}_n(m)=I_m(x_{n-B}(m)\cdots x_{n+F}(m)) \quad (1)$$

上式在估计丢失数据时，使用了其前 B 个特征和后 F 个特征信息。需要注意的是，对实时性的操作，F 要尽可能地小。

多项式插值的方法有很多，一般使用拉格朗日插值，对 $N+1$ 个特征矢量中的第 m 维，其插值形式为：

$$P_N(t)=L_0(t)x_0(m)+L_1(t)x_1(m)+\cdots+L_N(t)x_N(m) \quad (2)$$

其中拉格朗日系数 $L_n(t)$ 是 N 阶多项式。

一般为简化计算取一阶拉格朗日多项式，这样有：

$$\hat{x}_n(m)=\frac{t_n-t_q}{t_p-t_q}x_p(m)+\frac{t_n-t_p}{t_q-t_p}x_q(m) \quad (3)$$

其中，$\hat{x}_n(m)$ 是丢失的第 n 个特征矢量的第 m 维参数的估计，$p<n<q$，$x_p(m)$ 和 $x_q(m)$ 分别是 n 前后两个特征矢量的第 m 维参数。图5给出了这种插值的情况。

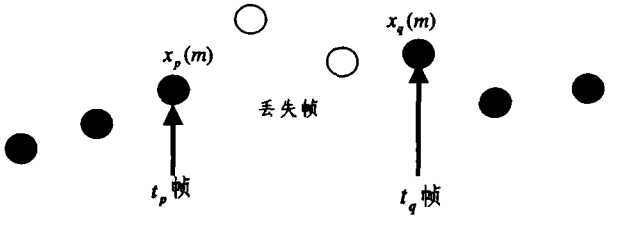

图5 丢失特征帧的多项式插值示意图

以上我们讨论了网络环境下由于语音编码和丢包所引起的语音识别性能下降的情况，以及相应的解决方法。从总体上看，本领域的研究工作才刚刚起步，还有许多问题亟待解决。

结束语 网络环境下的语音识别是近年来随着网络技术的发展而出现的一个新的研究课题。国际上该方面的研究已经开展起来，国内这方面的工作还比较少。相信不久的将来一定会有越来越多的人从事该方面的研究，也必将会有所突破。

参 考 文 献

1 Jim Van S，Jeff Z M. Investigation of Speech Recognition over IP Channels，CASSP 2002，IEEE Press. 3812～3815

2 Pelaez-Moreno C，Gallardo-Antolin A，Diza-de-Maria F. Recognizing Voice Over IP：A Robust Front-End for Speech Recognition on the World Wide Web. IEEE Transactions on Multimedia，2001，3(2)：209～218

3 Miner B. Robust Voice Recognition over IP and Mobile Networks. In：Proc. of the Alliance Engineering Symposium，2000. 1197～1200

4 Miner B，Semnani S. Robust Speech Recognition over IP Networks. In Akansu A N. ICASSP 2000，Istanbul，Turkey：IEEE Press. 1791～1794

5 Miner B，et al. Bobust Distributed Speech Recognition Across IP Networks. In：Proc. IEE Colloquium ISDS，1999. 6/1-6/6

6 Quercia D，Docio-Fernandez L，Garcia-Mateo C，Farinetti L，De Martin J C. Performance Analysis of Distributed Speech Recognition Over IP Networks on the AURORA Database，ICASSP 2002，IEEE Press. 3820～3823

7 韩纪庆，张磊，吕成国，王承发. 可穿戴计算机中的语音处理技术. 计算机科学，2002，29(5)：107～109