基于流媒体的在线语音合成系统

杨鸿武[1]，陶建华[2]，蔡莲红[2]

[1](西北师范大学 物理与电子工程学院，甘肃 兰州 730070)
[2](清华大学 计算机系 媒体所，北京 100084)

摘 要：利用流媒体技术实现了一个在线语音合成系统，能在 Internet/Intranet 上提供实时的在线语音合成服务. 系统采用了自适应技术适应网络状况的变化，并利用动态带宽分配技术管理客户端连接，保证了客户端连接的合成语音质量. 系统可应用于语音邮件、语音校对、远程教学等领域

关 键 词：TTS；语音合成；流媒体；网络带宽
中图分类号：TP393.02　　**文献标识码**：A　　**文 章 编 号**：1000-1220(2003)11-2000-04

An Online TTS System by Using Streaming Technology

YANG Hong-wu[1], TAO Jian-hua[2], CAI Lian-hong[2]

[1](*School of Physics and Electronics Engineering, Northwest Normal University, Lanzhou* 730070, *China*)
[2](*Department of Computer Science & Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: By using streaming technology, an online TTS system that can provide real time online TTS service on Internet/Intranet are realized in the paper. For Guarantee the quality of synthesized voice, the system can adapt the change of network status, and manage client link by using dynamic bandwidth allocation technology. The system can be used in voice mail, voice revision and remote education system.

Key words: TTS; stream media; multimedia; network bandwidth

1 引 言

计算机语音合成系统中，一般都带有一个几百兆的语音库，以获得自然度较高的合成语音. 一个部门可能有好几台计算机需要用到 TTS 系统，如果在每一台计算机中都安装一套 TTS 系统，必然要占用大量的系统资源. 可以考虑利用客户/服务器或浏览器/服务器结构实现分布式的语音合成系统. 虽然目前有很多技术均可以实现客户/服务器或浏览器/服务器结构的语音合成系统，但它们通常均要求终端用户将合成的音频文件下载到本地的计算机，然后再利用浏览器中的播放器插件或专门的媒体播放器来播放. 这种方法带来了两个突出的问题. 首先，由于必须下载音频文件，而音频文件的数据量通常都很大，在目前 Internet 普通用户接入速率较低的情况下，一句很短的合成文本的音频文件可能都需要很长的下载时间. 其次，由于必须将音频文件下载到本地计算机后才能播放，这必然占用本地计算机的存储资源，而且不能实现实时语音播放. 尤为重要的是，这样的方式往往导致音频文件的泄漏，难以控制资源的流失.

目前出现的流式媒体传输技术实现了声音、影像等多媒体信息的连续、实时传送，用户不必等到整个文件全部下载完毕，而只需经过几秒或数十秒的启动延时即可进行观看. 当多媒体信息在客户机上播放时，文件的剩余部分将在后台从服务器继续下载. 这不仅使启动时间大为缩短，而且不需要太大的缓存容量. 流式媒体传输避免了用户必须等待整个文件全部下载完才能观看的缺点，并且能适应多种网络带宽. 利用流媒体技术，我们实现了一个 Internet/Intranet 上的在线语音合成系统，客户端将待合成的文本提交给服务器端，由服务器完成文本的语音合成，并将合成语音以语音流的形式实时传送给客户端，解决了在线语音合成中音频文件的下载及数据量大的问题，可以于分布式的语音邮件、语音网页、语音校对、人机对话、远程教学等方面.

2 TTS 系统原理

TTS 系统的主要功能是将计算机中任意出现的文字，转换成自然流畅的语音输出. 它使得计算机不仅能够处理数据，显示图像和文字，还能像人一样的说话，从而使得计算机变得更为亲切、自然. 计算机语音合成技术经历了一个飞速发展的过程，目前，已经较为成熟并已大量应用在不同场合，如主页和电子邮件的阅读、文稿校对、人机对话、信息查询等等.

一般认为，语音合成系统包括三个主要的组成部份：文本分析模块、韵律生成模块和声学模块. 文本分析的主要功能是使计算机从这些文本中能够认识文字，从而知道要发什么音、怎么发音，并将发音的方式告诉计算机，另外还要让计算机知道文本中，哪些是词，哪些是短语、句子，发音时到哪应该停

收稿日期：2001-12-11　作者简介：杨鸿武，硕士，讲师，主要从事计算机语音合成方面的研究

顿, 停顿多长等等. 韵律生成模块决定最终系统能够用来进行

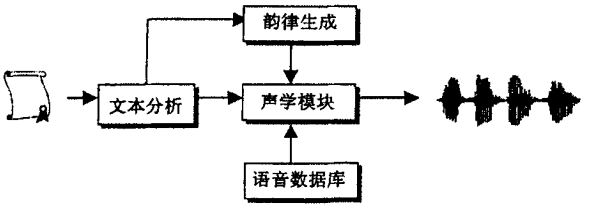

图 1　TTS 的系统框图

Fig. 1　Framework of TTS system

声信号合成的具体韵律参数, 如: 基频、音长、音强等. 声学模块具体实现合成语音的输出. 其系统结构如图 1 所示.

3　流传输技术

3.1　流媒体技术原理

流媒体是一种可以使音频、视频和其它多媒体能在 Internet 及 Intranet 上以实时的、无需下载等待的方式进行播放的技术. 流式传输方式是将多媒体数据经过特殊的压缩方式编码成一个个压缩包, 由媒体服务器向用户计算机连续、实时传送. 在采用流式传输方式的系统中, 用户不必像非流式播放那样等到整个文件全部下载完毕后才能看到当中的内容, 而是只需经过几秒或几十秒的启动延时即可在用户的计算机上利用播放器对流式文件解压后播放, 多媒体文件的剩余部分将在后台继续下载. 流式传输方式具有启动延时短、对系统缓存容量的需求低等特点, 其结构如图 2 所示.

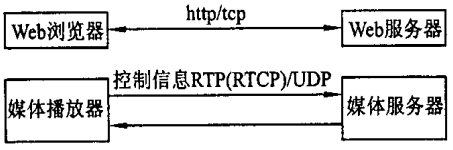

图 2　流媒体传输原理

Fig. 2　Theory of streaming media transportation

用户选择某一流媒体服务后, Web 浏览器与 Web 服务器之间使用 HTTP/TCP 交换控制信息, 以便把需要传输的实时数据从原始信息中检索出来; 然后客户机上的 Web 浏览器启动媒体播放器, 使用 HTTP 从 Web 服务器检索相关参数对播放器初始化. 这些参数包括目录信息、数据编码类型或媒体服务器地址. 媒体播放器与媒体服务器通过实时传输控制协议(RTCP)交换传输所需的控制信息. 媒体服务器使用 RTP/UDP 协议将流数据传输给媒体播放器, 一旦流数据抵达客户端, 媒体播放器即可播放输出.

3.2　流传输网络协议

目前, 支持流媒体传输的协议主要有实时传输协议 RTP (Real-time Transport Protocol)和实时传输控制协议 RTCP (Real-time Transport Control Protocol), 并通过多用途 Internet 邮件扩展 MIME 协议(Multipurpose Internet Mail Extensions)对流媒体类型进行识别.

RTP 被定义为在一对一或一对多的传输情况下工作, 其目的是提供时间信息和实现流同步. RTP 通常使用 UDP 来传送数据, 但 RTP 也可以在 TCP 或 ATM 等其他协议之上工作. 当应用程序开始一个 RTP 会话时将使用两个端口: 一个给 RTP, 一个给 RTCP. RTP 本身并不能为按顺序传送数据包提供可靠的传送机制, 也不提供流量控制或拥塞控制, 它依靠 RTCP 提供这些服务. 和 RTP 一起提供流量控制和拥塞控制服务. 在 RTP 会话期间, 各参与者周期性地传送 RTCP 包. RTCP 包中含有已发送的数据包的数量、丢失的数据包的数量等统计资料, 因此, 服务器可以利用这些信息动态地改变传输速率, 甚至改变有效载荷类型. RTP 和 RTCP 配合使用, 它们能以有效的反馈和最小的开销使传输效率最佳化, 因而特别适合传送网上的实时数据.

3.3　网络状况的自适应

多媒体应用会话持续时间一般比较长. 此期间客户端和网络状况部可能发生较大的变化, 媒体服务器应能自动适应这种变化. 网络带宽自适应是指媒体服务器根据客户端发送的 RTCP 包, 计算报文丢失、确认客户端看到的网络状况, 然后根据当前网络状况动态调节网络带宽. 网络带宽的调整基于 RTP 接收报告中的反馈信息, 包括用于计算报文丢失和报文延时抖动的信息. 媒体服务器根据反馈信息评价报文丢失及确认网络状况, 并相应地调节带宽.

媒体服务器收到 RTP 接收报告时, 首先分析所有接收成员报告并计算报文丢失, 延迟抖动和往返时间等, 然后根据报文丢失率将客户端观察到的网络状况分为拥塞、满载和轻载三类, 最后根据对网络状况的分析调整网络带宽.

网络状况变化后, 调节带宽的策略是: 在拥塞的情况下迅速减小带宽; 在连续观测到网络轻载时, 逐步增加带宽. 一般采取改变采样质量或编码方式等方法调节带宽. 比如说, 以 11.025kHz、16 位双声道采样数据为例, 占用带宽为 11025 * 16 * 2bit/s. 当网络处于拥塞时, 采用单声道采样, 则带宽降为 11025 * 16 * 1bits/s. 这样, 接收者尽管感到音频质量下降, 但不会明显地觉察到抖动或加大的延时.

3.4　服务器网络带宽的动态分配

当多个客户端连接到媒体服务器时, 服务器带宽被每个客户端共享, 每个客户端连接都要都要抢占媒体服务器的网络带宽, 因此服务器必须对带宽的共享进行控制, 以保证每个客户端连接都能获得最小的请求速率, 同时, 在还有可用带宽的情况下, 尽可能增加客户端连接的可用带宽, 使客户端连接获得最佳质量. 我们采用一种基于权重的最小- 最大带宽分配策略来分配每个客户端连接的带宽.

设服务器链路的带宽为 C, 客户端连接数为 S, 每个客户端连接的最小请求速率为 MCR, 峰值请求速率为 PCR, 每个客户端在建立连接时分配一个权重. 为了保证每个客户端连接都能获得最小请求速率, 则应有:

$$\sum_{i=1}^{S} MCR_i < C \tag{1}$$

当(1)式不能成立时, 服务器拒绝客户端连接请求. 否则, 服务器按以下算法给每个客户端连接分配带宽:

1. 每个连接入队列;
2. 给每个连接分配带宽 MCR;

3 给每个连接增加 CDP 的带宽, P 为增加比例

4 当某个连接的带宽达到 PCR 时, 该连接出队列, 当链路饱和时, 所有连接出队列

5 如果队列空, 则算法结束, 否则转 3

当链路饱和时, 如果有新的连接请求, 则按以下算法给新连接分配带宽:

1 给每个带宽大于 MCR 的连接减少 CDP 的带宽, 直到链路能给新连接分配带宽 MCR, 或(1)式不满足

2 如果(1)式不满足, 则拒绝新连接请求, 否则, 按分配算法分配带宽

3.5 流媒体的播放方式

在流媒体传输中, 从客户端与服务器端建立连接的方式划分, 流媒体的播放方式分为单播与组播. 单播是客户端与服务器之间的点到点连接, 从一台服务器送出的每个数据包只能传送给一个客户机. 组播是利用 IP 组播技术传递流媒体, 网络中的所有客户端共享同一流. 以这种方式传输的最大好处是可以节省网络带宽.

从客户端与服务器端的连接关系划分, 流媒体的播放方式分为点播与广播. 点播是客户端与服务器之间的主动的连接. 点播连接提供了对流的最大控制, 用户可以开始、停止、后退、快进或暂停流. 但这种方式由于每个客户端各自连接服务器, 所以会迅速用完网络带宽. 广播是指是用户被动接收流在广播过程中, 客户端接收流, 但不能控制流.

4 在线语音合成系统的设计与实现

4.1 系统的结构

在线语音合成系统由客户层、中间层和服务层组成, 其结构如图 3 所示.

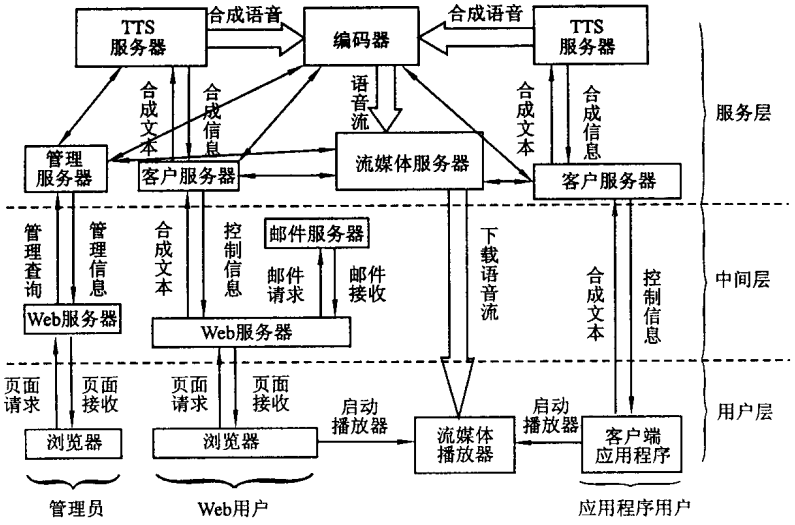

图 3 系统组成框图

Fig. 3 Architecture of online TTS system

服务层由 TTS 服务器、客户服务器、流媒体服务器、编码器和管理服务器组成.

TTS 服务器负责文本的语音合成, 并将合成的结果以实况源的形式发送给编码器.

编码器负责将合成语音编码成能适应不同网络带宽的流媒体文件格式. 编码器还通过在流中插入适当的同步标记, 实现合成文本与合成语音的同步输出.

流媒体服务器动态监测用户的网络带宽, 并根据用户的网络带宽将相应数据速率的语音流传送给客户端. 此外, 流媒体服务器还对客户端连接的带宽进行控制.

客户服务器接收客户端的合成请求, 对客户端的合成文本排队, 将合成文本发送给 TTS 服务器, 并向客户端发送控制信息, 以便让客户端与流媒体服务器建立连接, 使客户端接收合成语音. 为了减少用户的等待时间, 对于每个客户, 客户服务器并不是将该客户的全部文本交给 TTS 服务器合成, 而是每次从队列中取出一段要合成的文本传送给 TTS 服务器, 然后为下一个客户服务. 当某个客户的所有文本都合成完后, 将该客户从队列中删除.

管理服务器对整个系统进行管理和维护, 并生成日志文件.

根据应用模式的不同, 系统的用户分为三类: 第一类为管理员用户, 利用管理服务器对整个系统进行管理维护, 包括 TTS 服务器的配置与初始化、编码器的配置、流媒体服务器的配置、用户访问点的创建与维护、用户数量与带宽的限制以及系统日志的维护. 第二类为 Web 用户, 通过 Web 网页来提交合成文本. 用户可以通过网页提交一段合成文本, 也可以提交一个要合成的网页网址, 还可以提交电子邮箱帐号. Web 服务器获得提交信息后, 分别将文本内容、合成网页或者电子邮箱中的电子邮件取出来, 并将内容分别传送给客户服务器和客户端的 Web 浏览器, Web 浏览器通过启动客户端的流媒体播放器来收听合成语音. 第三类用户为应用程序用户, 利用客户端应用程序, 将要合成的文本传送给客户服务器, 当客

户端应用程序得到客户服务器的控制信息后, 通过启动流媒体播放器来收听合成语音.

系统的中间层由Web 服务器和邮件服务器构成. Web 服务器将服务层和客户层隔离开, 并且负责接收用户请求、转发命令和结果, 这在 Internet 上提高了系统的安全性.

系统的客户层由Web 浏览器、流媒体播放器和客户端应用程序组成. 流媒体播放器用来播放合成语音流, Web 浏览器和客户端应用程序用来接收用户请求、启动流媒体播放器和显示结果. 客户端应用程序直接与客户服务器交互, 一般应用在语音校对等客户/服务器模型中.

5 结　论

基于以上的设计, 我们在Windows 2000 平台上, 利用清华大学 SinoSonic 系统、Windows Media 服务、IIS 以及 DCOM 技术实现了一个 Internet/Intranet 上的在线语音合成系统. 系统可应用于语音邮件、语音校对、语音网页、远程教育, 并能给其它系统提供在线语音服务.

References:

1 Cai Lianhong, Wei Huawu. Research an Implementation of Text-to-Speech for Chinese [C]. ISSIPNN '94, Hong Kong, 1994. 4, 583～586.

2 Tao Jian-hua, Cai Lian-hong, Zhong Yu-zuo. The context-based method of creating chinese prosodic model [C]. ISSPR '98, 1998: 271～276.

3 Windows Media Technology. Microsoft Company [EB/OL]. http://www.microsoft.com/windowsmedia, 1999, 12

4 Yi-wei HOU Thomas. On network bandwidth sharing. For transporting rate-adaptive packet video using feedback [EB/OL]. Microsoft company. http://www.microsoft.com/china/research/downloads.