实用语音识别的场景标记辅助系统

杨庆涛[1], 李 昕[1,2], 郑 宇[3], 张 芸[1]

(1. 上海大学机电工程与自动化学院, 上海 200072; 2. 南京大学电子科学与工程系, 南京 210093;
3. 上海大学计算机学院, 上海 200072)

摘要: 标引是通过给音频-视频数据加入标记, 对其内容进行描述, 以便于信息的检索和查询。语音标引在媒体资产管理中扮演了很重要的脚色。介绍了一种基于 EBF 网络的语音标引辅助系统, 该系统可自动识别标引员所说的短语, 辅助标引员在视频媒体上实现标引。系统从语句中将这些短语分割出来, 并通过 EBF 神经网络进行建模。实验结果证明, 该系统具有实用性, 在媒体资产管理方面有良好的应用前景。

关键词: 媒体资产管理; 语音标引; EBF 网络

中图分类号: TP391 **文献标识码**: A **文章编号**: 1000-3630(2006)-05-0478-04

Multimedia scene labeling based on speech

YANG Qing-tao[1], LI Xin[1,2], ZHENG Yu[3], ZHANG Yun[1]

(1. School of Electromechanical Engineering and Automation, Shanghai University, Shanghai, 200072, China;
2. Department of Electronics Sciences & Engineering, Nanjing University, Nanjing 210093, China;
3. School of Computer Science and Technology, Shanghai University, Shanghai, 200072, China)

Abstract: The main objective of the indexing process is to assign labels to the audio-visual data in order to describe its content. Audio indexing plays a key role in this process. In this paper, a speech-based man-machine labeling system for media asset management is presented. The system recognizes the phrases spoken by the human annotator automatically and assists him to mark up shots of subjects in a video sequence. The phrases are segmented from short sentences and modeled by the elliptical basis function (EBF) networks. Experimental results indicate that the speech-based labeling system is practical and has great promise for media asset management.

Key words: media asset management; speech-based label; EBF neural network.

1 引 言

随着电视技术全面转向数字化、网络化, 信息技术亦逐渐成为广电行业主导技术。信息技术的应用可使媒体内容作为一种资产进行管理、挖掘和再利用。同时, Internet 网络和宽带网的普及, 更使媒体资产管理(MAM)系统的建立变得日益重要。

由于信息产业突飞猛进的发展, 文本、声音和图像等电子信息的发布量也呈几何级数增长。在当今社会中, 信息数据交换频繁, 每人每天都要接触大量信息, 面对大量的数据信息, 如何用一种简单、快捷的方法, 让用户能快速、准确地找到需要的内容, 就成为信息管理领域急需解决的难题。目前通过互联网或数据库针对文本搜索的功能已经广泛应用, 而基于声音或图像的检索还处于实验阶段。

对于电视台来说, 节目和素材是非常重要的资产。为了便于管理, 需要将节目分类并且对视频内容

收稿日期: 2005-10-13; 修回日期: 2006-01-06
基金项目: 上海市教委青年基金(04AB72), 上海市科委启明星计划资助(04QMX1441)。
作者简介: 杨庆涛(1978-), 男, 江苏人, 硕士研究生, 研究方向: 语音识别、信息检索。

进行标引。标引就是通过给多媒体数据加入标记,对其内容进行描述,以便于信息的检索和查询。语音标引在媒体资产管理中扮演了很重要的角色。语音标引的研究在国内目前基本还是空白,国外则偏重于将图像识别和语音识别结合在一起进行自动标引,但标引的效果不很理想[1]。因为目前自动标引的技术只能在媒体流中识别出事件的发生,至于是何种事件则无法判断[2]。例如在足球比赛中,可以通过识别观众的欢呼声检测出精彩场面,但是该精彩场面属于什么类别却无法判断,而解说员的声音受现场环境的干扰较大,直接影响到识别的效果[3]。因此,采用标引员和自动语音识别相结合的办法是比较实际的。本文所提出的就是一种基于神经网络的语音标引辅助系统。和现场解说员不同,标引员可以在相对比较安静的环境中,口述关键场面的内容,由自动语音识别系统进行识别并加以标记。

2　标引系统简介

以电视台的体育节目素材库为例,每年都有大量的体育节目播出,经过多年的积累,存储的体育节目素材的数量是惊人的。它们大都以磁带记录的模拟信号的方式保存,这样既占用空间,又因为磁带本身不适合长期存放,如遇潮湿、高温等,都将会导致画面质量的损失,尤其是一些珍贵的历史镜头或重要的节目片段,会因为保存不力而使素材无法使用。随着信息技术的不断发展,电视台节目制作和播出正在走向数字化和网络化,如何将这些节目素材有效地进行数字化存储管理和充分再利用是各个电视台所急待解决的问题。

存储的问题一般通过将各种模拟的节目素材转化成数字信号, 并使之能够支持多种的视频格式(DV25、MPEG4、REAL 等等)来解决。而要使节目素材能够被随时调用和检索,则需要通过标引的方式,对节目内容打上一定的标记。标引工作目前基本上都是依赖于人工,工作量非常大且因人为失误导致误标引。因此,加入智能化的手段,使人工标引转化为自动标引,是未来研究和发展的方向。将人工标引与智能化自动标引相结合,以内容为基础,提供多级智能检索方式,可与硬盘、非线性编辑以及播出系统无缝连接。基于语音的自动标引系统可以帮助标引员将口述的媒体内容记录下来,比如,标引员在标引一场足球比赛时会使用一些预先定义好的常用词语,如"某人进球"、"某人越位"、"某人犯规"等,系统将这些常用词语自动识别并记录下来,在指定的画面帧上进行标引。在检索时,用户的输入查询进入系统后将在已经建立的关键词特征库中进行,找出相关内容的画面帧。具体的标引流程如图 1 所示:

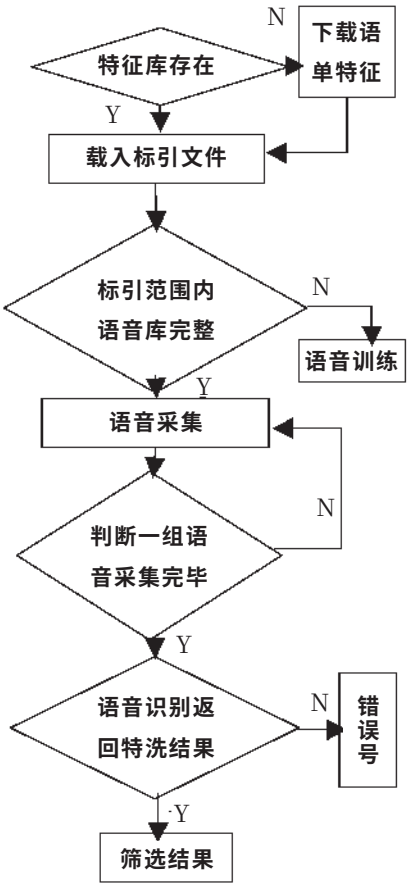

图 1　语音标引流程图

Fig.1　Process of speech label

3　语音识别的基本方法

输入的语音经过模数转换器成为数字信号后,可以用许多不同的方法来提取语音信号的特征参数。例如, 利用时域方法提取语音信号短时能量函数、平均幅度函数、短时平均过零率、短时平均幅度差函数以及短时自相关函数等;或者采用同态分析方法,提取同态分析的各种参数值,也可以利用线性预测分析方法提取线性预测编码(LPC)的各种参数或者 LPC 残留误差值等。为了识别不同幅度不同持续时间的语音,就需要对原始分析参数进行归一化处理,经过归一化处理后,数据格式可以有比较固定

的模式，而且可以进一步压缩。在样本模式存储器中存储有各种标准的语音模式参数，将标准的语音模式参数与待识别的语音参数进行比较，计算失真测度，按照预定的各种准则进行判别，最后就可以将判别结果显示出来。如图 2 所示。

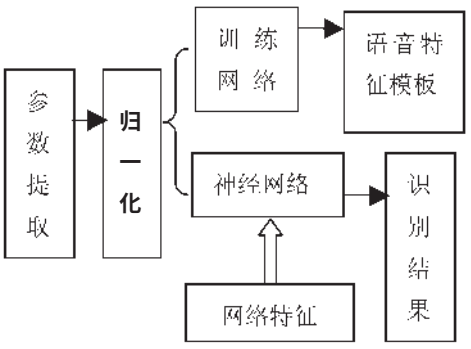

图 2 语音识别一般框图

Fig.2 Process of speech recognition

4 基于 EBF 网络的语音标引辅助系统

端点检测是实时语音识别系统中一个重要组成部分。本系统采用能量、过零率和相关法相结合的方法进行端点检测。过零率和相关法各有所长，相关法利用了语音的周期性，通过对自相关函数的衰减分析来区分语音和噪音。过零率对声母明显的语音非常有效。针对一些清音和浊音相分离的现象需要将两者结合起来使用。

分帧后的语音经过函数 $1-0.95z^{-1}$ 预加重和 10ms 的汉明窗后，提取 12 维 LPCC 参数，然后根据 LPCC 生成同维的 MFCC 特征[4]。

通常可以用隐马尔科夫(HMM)模型，矢量量化(VQ)或人工神经网络(ANN)的方法来对语音进行建模。在本项目中，考虑到训练与识别所使用的语音文本顺序信息并不一定相同，该语音信号短时顺序信息即 HMM 中的状态转移概率，对识别率的提高贡献不大。另外，矢量量化方法中，聚类的矢量仅用一个中心来表示，对语音特征的描述欠细致。因此，采用神经网络建模的方法为好。

本文中采用 EBF 神经网络[5]来构建语音帧的模型，EBF 网络是一种局部逼近网络，它能以任意精度逼近任一连续函数。一个 n 输入，m 维输出的 EBF 网络可以实现 n 维空间到 m 维空间的多维非线性映射。EBF 网络中对应着输入向量 x_p 的第 k 个输出的形式为：

$$w_{k0}+\sum_{j=1}^{J} w_{kj}\exp\{-\frac{1}{2\gamma_j}(x_p-u_j)^T\Sigma_j^{-1}(x_p-u_j)\}=y_k(x_p)\approx f(x)$$

$$p=1,\cdots,N \qquad k=1,\cdots,K$$

式中 u_j 和 Σ_j 分别为第 j 个基函数的均值矢量和协方差矩阵，w_{k0} 为偏置量，w_{kj} 是连接隐节点和输出节点之间的连接权值，是用来控制第 j 个基函数离散度的平滑参数。它可以由下式得出：

$$\gamma_j=\alpha\sum_{l=1}^{L}\|u_l-u_j)\| \qquad j=1,\cdots J$$

式中 u_l 表示在欧式空间中距离 u_j 最近的的向量，L 是这些向量的个数，α 则是控制基函数展度的系数。

EBF 网络的均值向量协方差矩阵通常可以通过下面三个步骤得出。首先，利用 K 均值法确定第 j 类训练矢量集的均值，中心点 u_j 由样本均值 $\hat{u}_j$ 估计得出：

$$u_j\approx\hat{u}_j=\frac{1}{N_j}\sum_{x\in\chi_j}x$$

这里，$x\in\chi$，如果满足 $\|x-\hat{u}_j\|<\|x-\hat{u}_j\|$，$\forall j\neq k$，$N_j$ 就是 χ_j 中的样本数，$\|\cdot\|$ 表示欧几里德范数。然后，由样本的协方差估计得到样本协方差矩阵：

$$\Sigma_j\approx\hat{\Sigma}_j=\frac{1}{N_j}\sum_{x\in\chi_j}(x-\hat{u}_j)(x-\hat{u}_j)^T$$

最后，通过最小二乘法得到输出权值 $\{w_{kj}\}$。

选取 24 维 LPCC 特征(12 维 LPCC+12 维差分 LPCC)以及 24 维的 MFCC 特征(12 维 MFCC+12 维差分 MFCC)作为输入矢量，实验中发现虽然采用 MFCC 特征在总体识别率上较 LPC 特征有较大幅度的提高，但在某些单词(如传球、防守)上采用 LPCC 特征的识别效果却优于 MFCC 特征。使用 24 维的 MFCC 和 LPCC 相结合的方法(12 维 MFCC+12 维 LPCC)在识别率上有所提高，对于号码和动作识别率分别可以达到 98%和 89.3%。

但上述三种做法虽然特征维数较高，程序运算时间消耗却较大，所以本文中采用一种排名加权融合的方法，即在一次判决中将低维 LPCC 和 MFCC (本文中使用 12 维)的排名结果进行融合运算，可以有效地解决存在的问题。在该方法中，两种特征矢量输入所得到的识别结果按相似度排名分别为 R_{LPCC} 和 R_{MFCC}，计算其加权和，即 $\omega_{MFCC}R_{MFCC}+\omega_{LPCC}R_{LPCC}$，其中 ω_{MFCC} 和 ω_{LPCC} 为权重[6]。实验结果表明，采用这

种方法,在无需对算法改动很大的情况下能取得较好的效果。

5　实验结果

文中介绍的是一个针对足球比赛的语音标引辅助系统。标引员用普通话短句描述赛场中关键场面,短句中应包括球员的姓名或号码以及动作或事件的信息,如 10 号犯规、12 号进球等。自动标引系统可以识别短句中的关键词,以对该场景进行说明。词库中包括 20 个球员号码以及 30 个常用的比赛术语(动作或事件),这样组合可以构成描述 600 个不同场景的短句。语音人机界面及识别程序在VC++6.0 平台上运行,考虑到系统的实用性,识别结果给出前 5 项,供标引员默认或选择。从实际效果看,响应时间可以达到实时要求。识别结果如表 1 所示。 从实验结果可以看出,采用排名融合的方法不仅在识别率方面较前三种方法有提高,且在运算耗时上有大幅度降低, 因为该方法的耗时仅相当于低维 LPCC 和 MFCC 运算耗时的线性相加, 可以更好地满足标引系统的实时性要求。

6　结　论

本文介绍的一种实用的语音标引辅助系统,它可以将标引员所讲的描述场景的短句分割成关键词并利用 EBF 神经网络进行建模或识别。为避免使用高维特征所带来的运算量增加、实时性降低的问题,文中提出了一种排名融合算法,在一次判决中将低维 LPCC 和 MFCC 输入矢量的排名结果进行融合运算。实验表明,采用该方法可使标引系统的识别率和实时性均有明显的提高。作为 MAM 系统的重要组成部分,该语音标引系统集成了语音训练、采集和分析功能的应用平台,它较一般的语音识别系统增加了在广电这一专业领域内的实用功能,极大的提高了电视台标引员的工作效率和服务质量。同时它增强了原有的 MAM 系统的应用功能, 提高了资源利用效率,从而产生可观的经济效益,随着多媒体检索、媒体资产管理的发展,该语音标引辅助系统将具有广泛的应用潜力和实用价值。

表 1　关键词识别率

Table 1　Retrieval performance of key words

特征组合方法	12^{th} LPCC+12^{th} 差分 LPCC	12^{th}MFCC+12^{th} 差分 MFCC	12^{th} LPCC+ 12^{th}MFCC	排名融合
首选(号码)	50%	69.2%	70%	72.5%
前五选(号码)	96.4%	97.3%	98%	98.7%
首选(动作)	51.7%	65.7%	64%	67.7%
前五选(动作)	81.3%	84.6%	89.3%	91.3%

表 2　几种方法的耗时比较

Table 2　Time consuming of the 4 methods

特征组合方法	平均耗时/s
12^{th} LPCC+12^{th} 差分 LPCC	1.47
12^{th} MFCC+12^{th} 差分 MFCC	1.95
12^{th} MFCC+12^{th} LPCC	2.01
排名融合	1.33

参　考　文　献

[1] Han Mei, Gong Yihong. Baseball scene classification using multimedia features[A]. 2002 IEEE International Conference on Multimedia and Expo[C]. 2002, 1: 821-824.

[2] Li Baoxin, Pan Hao, Sezan I. A general framework for sports video summarization with its application to soccer [A]. 2003 IEEE International Conference on Acoustics [C]. Speech, and Signal Processing (ICASSP'03), 2003, 3: 169-172.

[3] Chang Y L, Chang W Zeng, I Kamel, et al, Integrated image and speech analysis for content-based video indexing[A]. Proceedings of the Third IEEE International Conference on 17-23 June 1996[C]. 1996, 306-313.

[4] 赵力. 语音信号处理[M]. 北京, 机械工业出版社, 2003,167-168.
ZHAO li. Process of Speech signal[M]. Beijing, China Machine Press, 2003, 167-168.

[5] Mak M W, Kung S Y. Estimation of elliptical basis function parameters by the EM algorithm with application to speaker verification[A]. IEEE Trans. on Neural Networks[C]. 2000, 11(4), 961-969.

[6] 庄越挺, 潘云鹤, 吴飞. 网上多媒体信息分析与检索[M]. 北京, 清华大学出版社, 2002, 241-248.
ZHUANG Yueting, PAN Yunhe, WU Fei. Web-based Multimedia Information Analysis and Retrieval[M]. Beijing, Tsinghua University. Press, 2002, 241-248.