全自动中文新闻字幕生成系统的设计与实现

郑李磊,谢 磊,芦咪咪,王晓暄,杨玉莲,张艳宁
(西北工业大学计算机学院 陕西省语音与图像信息处理重点实验室,陕西西安 710072)

摘 要: 本文设计与实现了一个全自动中文新闻字幕生成系统,输入为新闻视频,输出为视频对应的字幕文本.以《新闻联播》为语料,实现了音频提取、音频分类与切分、说话人识别、大词汇量连续语音识别、视频文件的播放和文本字幕的自动生成等多项功能.新闻字幕的自动生成,避免了繁重费时的人工字幕添加过程.实验表明,该系统识别率高,能够满足听障等特殊人群和特殊场合的电视新闻收视需求.

关键词: 语音识别;广播新闻抄本;音频分类;说话人识别;字幕生成
中图分类号: TP39 **文献标识码:** A **文章编号:** 0372-2112 (2011) 3A-0-0

An Automatic Caption Generator for Mandarin Broadcast News

ZHENG Li-lei, XIE Lei, LU Mi-mi, WANG Xiao-xuan, YANG Yu-lian, ZHANG Yan-ning
(*Shaanxi Provincial Key Laboratory of Speech & Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi* 710072, *China*)

Abstract: Automatic broadcast news transcription converts speech into text by a large vocabulary continuous speech recognizer (LVCSR). This technique is an important prerequisite to various tasks, e.g., structural segmentation, semantic access and content-based retrieval of broadcast news. In this paper, we develop an automatic caption generator (ACG) for Mandarin broadcast news. The system integrates various functions, i.e., audio extraction from video, audio type classification and segmentation, speaker recognition, LVCSR, caption generation and video control. Experiments show that the system can achieve high speech recognition accuracy. A potential deployment of ACG is to help the hearing impaired and elderly people in enjoying TV programs.

Key words: speech recognition; broadcast news transcription; audio classification; speaker recognition; caption generation

1 引言

电视广播新闻节目是人们获取信息、掌握国家方针政策、了解国内外重大事件和社会发展现状的最普遍渠道之一.在广播新闻观众中,一些特殊人群,如听障人群、听力弱化的老年人等,只能通过新闻视频和其标题字幕来揣测新闻的主旨,而无法从伴音中获取完整的语义信息.为新闻节目添加全程字幕是解决这一问题的首选办法.字幕可以帮助这些特殊人群更好地了解新闻内容,另外还能满足人们在一些特殊场合(如需要保持安静的医院和嘈杂的车站候车厅)的收视需求.

在欧美等发达国家,电视新闻节目已逐步配备字幕.我国香港特别行政区也基本实现了电视节目的字幕化工作.在《中国残疾人事业"十五"计划纲要》中,"电视新闻、影视剧要逐步加配字幕说明"更被列为重点实施的政策之一[1].目前,国内大部分的电视剧均已配备字幕,而配备字幕的新闻节目甚少,这是因为新闻节目具有高时效、更新快的特点,其数量每一天都在大幅度增加.要想为这些新闻节目手工制作字幕费时又费力.因此,发展自动化的字幕生成技术是满足当前新闻字幕需求的最有效途径.

实现字幕自动化生成,关键是要解决音频向文本转换(Speech to Text, STT)的问题.因此大词汇量连续语音识别(Large Vocabulary Continuous Speech Recognition, LVCSR)技术是全自动字幕生成系统的核心,高识别率则是系统走向实用的基本要求.目前,大词汇量连续语音识别开始走向商用化应用,包括 IBM ViaVoice 和 Dragon Dictation 在内的听写系统相继推出,广播新闻自动抄本(Broadcast News Transcription)系统[2]和基于语音识别的海量音频数据处理[3]成为研究热点.例如,Gills Bouliance 等人对法语电视新闻节目的字幕自动生成进行了研究[4],主要针对特定主题的新闻,实现了音频到文本

收稿日期:2010-06-22;修回日期:2010-08-08
基金项目:国家自然科学基金(No.60802085);国家高技术发展计划(2009AA01Z315);教育部科技创新工程重大项目(No.708085);教育部新世纪优秀人才支持计划;教育部高等学校博士点学科专项基金(No.20070699015);西北工业大学基础研究基金(No.W018103)

的转换.Akio Ando 等人设计了一个实时抄本系统,实现了日本广播新闻同步字幕的制作[5].这些系统侧重于大词汇量连续语音识别系统的正确率和识别速度,假设新闻音频仅包含语音成分,而没有关注字幕生成系统的完整性.一个完整的字幕系统应能完成音频提取、音频分类与切分、语音识别、字幕生成、字幕视频同步播放等一系列功能.因为新闻音频中通常伴有一些非语音成分,如片头音乐、广告,或背景音等,这些成分无需进行语音识别,全自动字幕生成系统应当在语音识别前将它们从新闻音频中甄别并分离出去.

本文设计了一个全自动中文广播新闻字幕生成系统,实现了从中央电视台《新闻联播》视频节目到字幕文本的一键式全自动制作,具有音频提取、音频分类与短句切分、说话人识别、语音识别、标准字幕生成与显示等功能.系统具有界面简洁、操作简单、语音识别率较高等特点,为特殊人群和在特殊场合观看电视节目提供了一条便捷的途径.

2 系统概述

本文的目标是实现中文新闻字幕的一键式全自动制作.实现这一目标需要完成的主要功能模块为音频提取、音频分类与切分、说话人识别、语音识别和字幕生成与显示.如图 1 所示,系统首先从视频文件中提取出音频流,然后将音频切分成语音与非语音片段,并将语音部分继续切割成便于识别的短句.接下来,识别每一个短句的说话人,根据建立的说话人专有声学模型,使用大词汇量连续语音识别系统对短句进行识别,得到相应的识别文本,即抄本 (transcripts).最后将识别文本转换成标准字幕,与视频同步播放.系统的输入为各种常用格式视频文件,如 wmv 文件、avi 文件、rmvb 文件等;系统的输出是 srt 标准文本字幕,可被暴风影音等常用视频播放软件自动加载.

四大模块的功能分别介绍如下:

(1)音频提取:本文采用 E M Total Video Converter 2.41 音频提取工具[6]从新闻视频中提取音频,该工具参数设置灵活,可实现多种音视频文件的相互转换.本文从视频中提取的音频属性为:wav 格式、单声道、16bit、采样率 16kHz.

(2)音频分类与切分:首先利用研究组前期设计开发的音频分类系统[11]对提取的音频进行分类,判断固定窗口内的音频片段的类别:静音、语音和非语音;然后根据得到的整段音频类别信息,设计切分算法从音频中截取出易于识别的语音单元.音频切分后得到以一定时长的静音作为分割边界的、语义相对完整的"类句子"单元.为方便起见,本文仍将这些类句子单元称为句子.

(3)说话人识别与语音识别:为提高识别率,本文采用特定人语音识别策略,首先对句子进行说话人身份判别,然后采用每个说话人的特定声学模型进行语音识别.对《新闻联播》中的常见说话人建立各自的声学模型,语言模型为词级三元文法(Trigram),由大量中文新闻文本语料训练而成.采用 Julius[9]作为大词汇量连续语音识别工具进行句子识别.本模块最终生成若干文本文件,存储着每一个句子的识别文本结果以及对应的起止时间信息.

(4)字幕生成与显示:设计字幕生成算法,根据语音识别模块得到的文本内容和对应时间信息,生成 srt 文本字幕.视频播放器(如暴风影音)自动加载该字幕并播放视频,实现字幕与语音的同步.

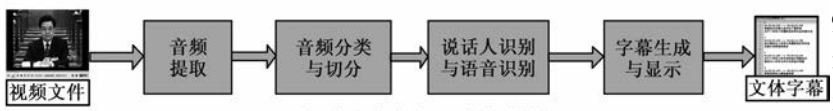


图1 全自动字幕生成系统结构框图

3 音频分类与切分

新闻节目的音频信息种类丰富,既有播音员、记者的语音播报,又有穿插的转场音乐、广告、背景音等多种音频形式.实现对音频媒体的有效分类和分割,是多媒体信息处理的重要前提.尤其对于针对广播音频的语音识别系统来说,音频分类是一个不可缺少的预处理过程.通过音频分类区分出音频中的语音部分和非语音部分,将不包含说话内容的非语音部分分离出去,并将语音部分切分成若干"类句子"单元.这些句子的长度适中,语义相对完整,易于被识别器识别.

3.1 音频分类

本文采用研究组前期设计的音频分类系统[11]完成广播音频的自动分类.该系统是一个集多种音频特征、多种分类器和多种分类决策方法为一体的音频自动分类系统.基于 Marsyas 工具包[10],提取 12 种时域、频域特征(如表 1 所示),可在高斯混合模型(Gaussian Mixture Model,GMM)、K-近邻(K-Nearest Neighbor,KNN)或支持矢量机(Support Vector Machine,SVM)之间进行分类器选择.可以设置诸如帧长、段长等配置参数.在针对 CCTV《新闻联播》音频分类测试中,经过经验实验,设定帧长为 32ms,段长为 0.16s,即一个音频段包含 5 个音频帧,每段做一次分类决策.该工具在核函数为径向基(RBF)的 SVM 分类器的情况下,对《新闻联播》音频中语音/非语音/静音的分类 F-measure 值分别能达到 92.88%、92.52%和 95.13%[11].图 2 为该分类系统对一段音频的分类结果.从整体上看,连续的语音是由句子和停顿交错构成的,停顿可作为音频切分的重要依据.

表 1　部分音频特征

特征名称	说明
频谱质心 (Spectral Centroid,SC)	频谱能量的集中点,一般来说,此值越小,说明越多的能量集中在低频范围内.
频谱差分幅度 (Spectral Flux,SF)	一个音频段中的相邻两帧之间谱的平均变化量.
频谱截止频率 (Spectral Roll off Frequency,SRF)	把频率小于等于该值的所有信号的能量相加,其和为总能量的固定比例(可设定).
频谱峰度 (Spectral Kurtosis,SK)	描述频率分布曲线形态陡缓程度的统计量.
美尔频率倒谱系数 (Mel-Frequency Cepstral Coefficients,MFCC)	在美尔频率尺度上进行频谱分析,美尔频率尺度与实际频率关系为 $Mel(f)=2595\lg(1+f/700)$.
线性预测倒谱系数 (Linear Predictive Cepstral Coefficients,LPCC)	由线性预测系数变换得到.线性预测的基本思想:用过去的 p 个样点值来预测未来的样点值.
短时能量均方值 (Root-Mean-Square,RMS)	一帧的短时能量的均方值.
过零率 (Zero-Crossing Rate,ZCR)	一帧中信号波形穿过横轴(零电平)的次数.
高过零帧比率 (High Zero-Crossing Rate Ratio,HZCRR)	一个音频段内过零率超过 zcr 值的帧数目,zcr 值为所有帧的过零率平均值的 1.5 倍.
低能帧比率 (Low Short-time Energy Ratio,LSTER)	一个音频段内能量低于此段内短时能量平均值 0.5 倍的帧数目
噪音帧比率 (Noise Frame Ratio,NFR)	一个音频段内噪音帧所占比例

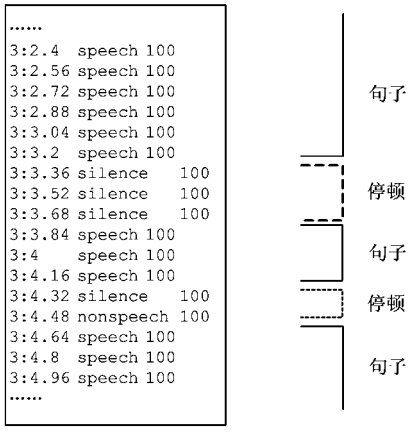


图2　音频片段分类结果

3.2　音频切分

在对音频段进行分类之后,需要根据语音停顿将语音流划分为易于进行识别的短句.检测句子的开始与结束是其中的关键,因为只有达到较高的端点检测精度,才可以做到有的放矢,实现对句子长短和数目的控制.本文设计的音频切分算法的基本策略是:以进入连续语音段之前的静音段或非语音段的时间点作为句子的开始时间,以结束连续语音段时的最后一个语音段的时间点作为句子的结束时间.

但是,仅凭上述方法检测句子端点会造成两种极端情况:一是有很多极短的句子,某些长度仅为一到两个音频段.这些句子通常只包含一两个词语,甚至不包含任何有效的语音信息;二是出现若干长句,某些长达数十秒甚至几十秒,包含有若干语义完整的单元.这两种情况都会严重影响识别率.

在算法中我们增加了三个变量来改善这两种情况:

(1)停顿最小长度限制:指一个停顿至少要包含的静音段或非语音段的个数.这个限制变量的作用是忽略较短的伴音信息,比如主持人的瞬时换气等,以保护一句话的完整性.本文设定的最小停顿为 2 个音频段,即连续语音单元中的单个非语音单元不会被视作为一个停顿.

(2)句子最小长度限制:指一个句子至少要包含的语音段的个数.这个限制变量的作用是滤除掉音频中的短时无效信息,比如主持人的轻咳等.本文设定的最小句子为 3 个音频段,即忽略总长小于 0.48 秒的语音单元.

(3)句子最大长度限制:从严格意义上讲,这不是一个限制变量,而是一个提示变量.当一个句子所包含的语音段的个数达到一定限度时,则采取方法使句子尽快地结束,例如,减小停顿最小长度限制,这样可以有效限制句子的长度,提高该句的识别准确率.本文设定的限度是 50 个音频段,达到这个限度以后即使是单个非语音单元也会被视作为一个停顿.

4　说话人识别

一个广播新闻节目的说话人通常包括一至两个播音员、众多记者、发言人和被采访者.某些说话人经常出现,例如国家领导人、播音员、著名的记者等.如果根据不同的说话人各自独有的发音特点对这些经常出现的说话人进行基于特定人的语音识别(Speaker Dependent Speech Recognition),那么就能有效地提高语音识别正确率.从连续的广播音频流中分割识别出不同说话人是进行特定人语音建模和语音识别的先决条件,只有先判断出句子的说话人身份,才能利用该说话人的特定声学模型进行语音识别.

4.1　识别策略

说话人识别模块的任务是对播音员以及一些新闻中常见的说话人(如国家领导人)建立自适应模型,对除此之外的说话人建立统一的模型,因此采用一种分层说话人识别策略.对某一待识别句子,首先进行 $N+1$ 的识别,其中 N 为已知身份的说话人模型,1 对应一个 unknown 模型.如果属于 N 个已知说话人之一,则直接进行标注;而如果被识别为 unknown 说话人,则继续

进行性别识别,根据识别结果,该说话人将被标注为 unknown 男性或 unknown 女性.本文采用的说话人识别系统为研究组前期设计的基于 Alize 工具包的说话人自动标注系统[12].

4.2 系统实现

本文使用的说话人识别系统以切分好的句子为处理对象,提取其美尔频率倒谱系数(Mel Frequency Cepstrum Coefficient,MFCC)作为用于识别的特征矢量序列.说话人模型采用高斯混合模型,基于 Alize 工具包编写建模和判别决策算法.

建模任务是为每个说话人建立一个高斯混合模型,整个过程包括模型初始化和训练两个阶段.初始模型的建立使用对应说话人的全部训练数据的 MFCC 特征,调用 Alize 中的相关函数,计算所有特征矢量的均值和协方差,用这两个值作为模型的初始均值和协方差矩阵,模型中各成员分布的初始权重相等.然后基于最大似然准则(Maximum Likelihood,ML),利用期望最大化(Expectation Maximization,EM)迭代算法训练模型,目的是使迭代出的模型能最好的代表说话人的语音特征.EM 算法使用初始模型的参数估计新的模型参数,新模型再作为当前参数进行训练,如此迭代直到模型收敛.系统参数可自主设置.通过经验实验,我们最终设定高斯混合分量数为 256、EM 迭代次数为 30、模型训练次数为 10.

识别结果的判定采用计算对数似然率(Log-Likelihood Ratio,LLR)的方法.在说话人识别中,假设有 S 个说话人,对应的 GMM 模型分别为 $\lambda_1,\lambda_2,\cdots,\lambda_S$,对于一个待识别序列 X,识别任务即为找到使之有最大后验概率的模型所对应的说话人 λ_S.假定每个说话人的出现为等概率事件,使用对数得分,识别的任务就是计算

$$\hat{S}=\arg\max_{1\leqslant k\leqslant S}\sum_{t=1}^{T}\log p(x_t|\lambda_k) \qquad (1)$$

S 即为检测出的说话人.在对 2.5 小时《新闻联播》音频的测试中,说话人识别与标注系统的误识率为 0.083.

5 大词汇量连续语音识别

大词汇量连续语音识别是字幕生成系统的核心,分为特征提取、模型训练、识别三大模块,如图 3 所示.本文采用基于隐马尔科夫模型(Hidden Markov Model,HMM[13])的识别策略,HMM 模型和语言模型的训练使用 HTK 工具包[7]完成,最后选用 Julius 工具包[9]作为识别器.

5.1 特征提取

将以句子为单元的语音片段加汉明窗分帧,帧长为 25ms,帧移为 10ms;提取 MFCC 特征及其一阶二阶差分,共计 39 维;由于中文是声调语言,声调的特点明显地反映在基频(F0)上,因此另采用 YIN pitch tracker[8]提取基频特征及其一阶二阶差分.最终的混合特征维数为 42 维.

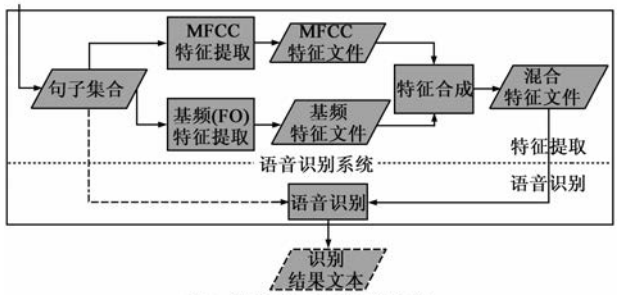


图3 语音识别系统流程图

5.2 声学建模

建模单位是考虑了中文声调和词内扩展的上下文相关的三音素 HMM 模型,模型生成示意如下:

```
计算机 j i4 s uan4 j i1
→ j-i4+s  i4-s+uan4  s-uan4+j  uan4-j+j1
```

每个三音素 HMM 模型都有三个有效状态,每个状态的混合高斯个数为 8.模型数量近 300K,而后经过决策树状态捆绑以减少参数.模型训练采用长达 56 个小时的音频数据,来自下列两个语料:

(1)Speech Lab in a Box 中文语音数据,时长达 31 小时有余;

(2)经过手工标注的 2007 年 7 月份到 2007 年 12 月份之间的 64 个《新闻联播》节目音频,总时长达 25 小时有余.

训练的声学模型分为非特定人模型和特定人模型两类,非特定人模型采用全部 56 个小时的音频数据训练而成,而后通过说话人自适应,为每个说话人(两个主要主持人、其他男说话人、其他女说话人)单独建立特定人模型.

5.3 语言建模

采用 HTK 中的语言模型建模工具建立词级三元文法(trigram).训练数据来自下面三个中文文本语料:

(1)自 2000 年 4 月至 2009 年 2 月的《人民日报》文本数据;

(2)自 2002 年 9 月至 2009 年 2 月的《新闻联播》音频和文本数据;

(3)自 2006 年 7 月至 2009 年 2 月的《新闻 30 分》音频和文本数据.

这些语料从相关网站下载,经过网页正文提取、文本格式化、正向最大匹配分词等步骤,产生最终约 524MB 的文本数据,用于三元文法的训练.本文所用中文字典的词条数目为 58386.

5.4 识别器及识别实验

大词汇量连续语音识别器采用日本京都大学开发

的 **Julius** 工具包[9].Julius 为双步(two-pass)识别器,支持三元语法,且同 HTK 文件格式兼容,性能优异.

对识别结果采用如下识别精度考量的方法:

$$Accuracy = \frac{N - D - S - I}{N} \times 100\% \qquad (2)$$

N:原始抄本(reference)文件中识别单元(如词)的个数;
D:对应于参考序列,识别结果中被删除的识别单元(如词)个数;
S:对应于参考序列,识别结果中被替换的识别单元(如词)个数;
I:对应于参考序列,识别结果中被插入的识别单元(如词)个数.

首先采用上述语言模型和非特定人声学模型,进行了非特定人语音识别实验,实验数据来自 2007 年《新闻联播》音频中随机选取的 609 个句子,来自演播室和室外等各种场景,且测试语料与训练语料不交叉.实验结果如表 2 所示,该语音识别系统词级的正确率达到 75.03%,字级、音节级以及基础音节级的正确率均在 80%以上.

表 2　非特定人语音识别实验结果

测试级	识别精度(Accuracy)%
词级	75.03
汉字级	80.86
音节级	81.81
基础音节级	83.69

而后我们进行了特定人语音识别,实验数据为 2007 年《新闻联播》音频中随机选取的邢质斌、张宏民和其他男、女说话人共计 843 个句子,且与训练语料不交叉,实验结果如表 3 所示.由表 3 可见,加入说话人自适应后的语音识别率有较大提高,词级平均正确率达到了 89.69%.

表 3　说话人自适应后语音识别实验结果(识别精度 %)

说话人类别	测试句子数	词级识别精度	汉字级识别精度	音节级识别精度	基础音节级识别精度
邢质斌(女)	50	95.08	96.86	96.86	97.43
张宏民(男)	53	86.59	91.87	91.76	92.56
其他女	340	89.62	94.13	94.60	95.49
其他男	400	87.46	91.84	92.18	93.24
平均	—	89.69	93.68	93.85	94.68

6　字幕生成与显示

字幕的种类有很多种,现在比较流行的字幕格式有图形格式和文本格式两类.相对于图形格式字幕而言,文本格式字幕有尺寸小、格式简单、便于制作和修改等特点.其中 srt 格式的文本字幕使用最为广泛,能兼容各种常用的媒体播放器,暴风影音、QQ 影音等均可自动加载该类型字幕.

6.1　字幕的生成

根据语音识别模块提供的句子文本信息和相应时间信息,设计字幕生成算法:按照一句时间代码 + 一句字幕的格式向 srt 文件写入文本.同时为了优化显示效果,方便观众观看字幕,将识别结果中较长的句子切分为多行显示.

6.2　字幕的显示

采用 VC、MFC 编程技术,实现视频播放器界面的设计和字幕的显示[14].系统内嵌暴风影音播放组件,可以自动加载生成的字幕,并在视频播放时同步显示.系统集安装使用于一体,考虑用户多方面的需求,共设计了九个模块,相互搭配完成各项功能.这九个模块是:文件操作、视频显示、视频控制、字幕制作、字幕加载、进度栏控制、菜单控制、快捷键控制、帮助等.

6.3　全自动字幕生成系统界面

图 4 为全自动字幕生成系统在完成字幕制作后加载字幕播放时的界面效果.该界面主要的特点有:一键字幕制作、内嵌功能强大的播放器、直观的进度栏设计、详尽的菜单、丰富的热键、友好的图形化界面、完备的帮助系统等.

图4　全自动字幕生成系统界面

7　结束语

本文设计了一个完整的全自动中文新闻字幕生成系统,实现了针对《新闻联播》节目的字幕一键式生成.该系统输入为《新闻联播》视频,输出为 srt 格式标准字幕文件,具有音频提取、音频分类与切分、说话人识别、大词汇量连续语音识别、srt 文本字幕生成、视频播放器等多个模块.系统界面设计简洁、操作简单,语音识别率高.

下一步的工作为:

(1)设计精细化音频分类工具,将音频分类为纯语音、带噪语音、带音乐语音、音乐、静音等更为精细的音频类型,随之改进音频切分算法截取更符合实际情况的合理语音单元.

(2)本文所设计的大词汇量连续语音识别系统的识别率仍有提高空间.拟采取加大训练数据、为更多的说话人建模等方法以提高识别率,进一步提高全自动字幕生成系统的实用性.

参考文献

[1] 中国残疾人事业“十五”计划纲要(2001 年 – 2005 年)[OL]. http://www.gov.cn/gongbao/content/2001/content_60808.htm,2001-04-10.

[2] M J F Gales, D Y Kim, P C Woodland, D Mrva, R Sinha, S E Tranter. Progress in the CU-HTK broadcast news transcription system [J]. IEEE Transactions on Speech and Audio Processing, 2006, 14(5): 1513 – 1525.

[3] C Chelba, T J Hazen, M Saraclar. Retrieval and browsing of spoken content [J]. IEEE Signal Processing Magazine, 2008, 25(3): 39 – 49.

[4] G Boulianne, J F Beaumont, M Boisvert, J Brousseau, P Cardinal, C Chapdelaine, M Comeau, P Ouellet, F Osterrath. Computer-assisted closed-captioning of live TV broadcasts in French [A]. Proceedings of INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing[C]. United Kingdom: DUMMY PUBID, 2006. 273 – 276.

[5] Akio Ando, Toru Imai, Akio Kobayashi, Haruo Isono and Katsumi Nakabayashi. real-time transcription system for simultaneous subtitling of Japanese broadcast news programs[J]. IEEE Transactions on Broadcasting, 2000, 46(3): 189 – 196.

[6] E M. Total Video Converter Command Line Version 2.44 [OL]. http://www.effectmatrix.com/total-video-converter-command-line/index.htm, 2009 – 2010.

[7] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev and Phil Woodland. The HTK Book (for HTK Version 3.4)[M]. United Kingdom: Cambridge University Engineering Department, 2006.

[8] A de Cheveigne, H Kawahara. YIN, a fundamental frequency estimator for speech and music[J]. Journal of the Acoustic Society of America, 2002, 111(4): 1917 – 1930.

[9] Tatsuya Kawahara and Akinobo Lee. Multipurpose Large Vocabulary Continuous Speech Recognition Engine[M]. Japan: Nara Institute of Science and Technology, 2001.

[10] G Tzanetakis. Marsyas: Music analysis, retrieval and synthesis of audio signals MARSYAS[A]. Proceedings of the Seventeen ACM International Conference on Multi-Media [C]. New York: ACM, 2009. 931 – 932.

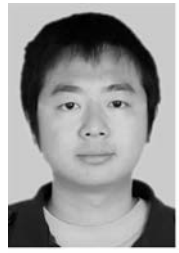
[11] 李中华.广播新闻音频分类:对比研究[D].陕西西安:西北工业大学,2008.

[12] 卢咪咪,谢磊,郑李磊,杨玉莲,张艳宁.基于 Alize 工具包的广播音频播音员自动标注系统[A].第 5 届和谐人机环境联合学术会议(HHME2009)论文集[C].陕西西安,2009.

[13] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257 – 286.

[14] 丁博.语音识别辅助工具的设计与开发[D].陕西西安:西北工业大学,2008.6.

作者简介



郑李磊 男,1989 年 9 月生于湖北仙桃.西北工业大学计算机应用专业硕士生,主要研究方向为音频、语音与语言处理.
E-mail: lzheng@nwpu-aslp.org

谢 磊 男,1976 年 10 月生于河北石家庄.西北工业大学计算机学院教授,教育部新世纪优秀人才.主要研究方向为音频、语音与语言处理,多媒体技术及人机交互.
E-mail: lxie@nwpu.edu.cn
网址: http://lxie.nwpu-aslp.org

张艳宁 女,1967 年 10 月生于陕西宝鸡.西北工业大学计算机学院教授,教育部新世纪优秀人才.主要研究方向为多媒体技术及人机交互、计算机视觉与模式识别.
E-mail: ynzhang@nwpu.edu.cn