跨信道说话人识别语音库的设计与研究

李琳，万丽虹，黄玲，洪青阳
(厦门大学信息科学与技术学院，厦门，361005)

文　摘：建立一个适合于特定说话人识别系统的汉语语音数据库，对推动说话人识别技术的研究和应用具有重要意义。本文针对说话人识别中信道差异问题，设计并建立了跨信道说话人识别语音库 XMU-MultiChannel。本文首先分析比较了国内外现有的说话人识别语音库，设计了跨信道说话人识别语音库的语料内容、采集系统、录音方案以及标注和存储方式，进一步基于这个数据库实现了 PLDA 参数更新的跨域迁移策略，以解决小样本信道 PLDA 建模的困难，并在语音样本有限的前提下提高了说话人识别系统的识别性能。

关键词：语音数据库；说话人识别；跨信道 ；跨域迁移
中图分类号：TN912.34

随着说话人识别技术的突飞猛进，目前在限定条件（特定环境/信道/人数）下的识别已经可以获得比较满意的结果。但是现有的说话人识别系统仍存在许多困难，例如：1）信道不匹配问题；2）背景噪声问题；3）短时语音识别问题；4）说话人情感、健康状况的变化对语音产生的影响等等。这些问题对说话人识别系统的性能干扰都非常显著。本文针对跨信道研究的需要，设计和建立跨信道说话人识别语音库。

跨信道问题是由于语音的传输过程中信道或者采集设备会对说话人的语音产生不同程度的畸变，在很多实际应用领域是无法避免的，如国防侦听、水声通信、司法鉴定的过程中，采集语音信息所使用的设备往往是不可控的；对于通话信道而言，模电转换、压缩编码、传输错误、丢包弃包等都会使语音产生变化，如当前国内的电话网络中包括 GSM、CDMA、TD-SCDMA 等通信协议，也有 PSTN、IP 电话等固定电话传输信道，还包括当下越来越多的类似 Skype、Google Voice 产品之类的互联网语音应用。不同的终端，不同的接入方式以及噪声等对声音尤其是说话人的正确识别产生影响[1]。要研究跨信道问题就需要大量语音数据，所以对跨信道说话人识别语音库建设的要求非常迫切。

1　说话人识别语音库的现状

国际上非常重视语音库的建设，均投入大量人力、财力建设相应的语音库。美国的LDC(Linguistic Data Consortium)，欧洲的ELRA(European Language Resources Association) 以及美国的OGI(Oregon Graduate Institute)[2,3]均是长期致力于语言相关资源的大规模开发和共享的主要组织和机构。

LDC于1992年建立，它开发的TIMIT语音库包含有630名美式英语说话人(男性438人/女性192人)，每人10句话，是最早应用于说话人识别研究的语音库之一。TIMIT同时包含有一系列经过二次处理的派生语音库，例如FFMTIMIT(较远距离麦克风录制)、NTIMIT(通过长途电话网传输后录制)、CTIMIT(通过移动通信网传输后录制)、HTIMIT(直接通过固定电话录制)等。LDC开发的Switchboard I-II语音库分别包含了543人和657人的电话对话，可用于说话人识别辨认和确认研究。此外，LDC还发布了第一个大容量、高质量的语音库YOHO语音库，专门用于政府门禁安全控制应用。LDC中还收录了KING语音数据库，King语音数据库是于1987年由国际电话和电报公司--ITT依据美国政府的研究合同收集整理[2]。1992年，对KING语音数据库的原始数据重新加工得到后来的KING-92语音数据库。语音数据库KING由51名男性说话者通过两种不同的信道采集而成：一个信道是各种各样的电话，另一个是高质量的麦克风。这51名说话者被进一步分成两组，并被安排在不同的位置录制。每个说话者通过每种信道采集10段语音，每段大约持续30秒到60秒的时间。

ELRA开发的SIVA电话语音库是专门针对意大利语说话人识别的，它通过公用电话交换网录制了671人(男性335人/女性336人)的超过2000次的电话语音[2,5]。Polyvar是由ELRA发布的一个说话人确认的法语电话语音数据库，说话人主要是瑞士人和法国人[2]。录音内容包括一些朗读和自然语音，大概有160个小时的语音。其中31个说话者进行了2~9次通话，41个说话者进行了10次以上的通话。Polycost语音数据库是ELRA在开展欧洲cost250项目中创建的，

用于说话人确认[2,6]，该语音库包含了134名欧洲人(男性74人/女性59人)的英语电话录音。

OGI开发的CSLU Speaker Recognition语音库是面向大容量文本相关/文本无关的说话人鉴别和确认应用[2,4]。在两年内采集了500名说话人，每人至少12次通话，分别在安静或嘈杂的环境中使用不同的电话进行通话，如无绳电话、移动电话、付费电话等。

此外，西班牙的Politenica大学建立了麦克风和电话双信道的西班牙语音库AHUMADA[3,7]。值得一提的是，该语音库特别考虑了导致说话人语音变化的自身因素（如年龄，说话方式，口音等）及外在因素（麦克风，传输信道，回声等），进行数据采集和语音文本的设计。

近年来，国内建成的语音语料库也很多。例如中国科技大学、中国科学院声学研究所、中国社会科学院语言研究所联合开发的汉语语音识别资料库; 中国社会科学院语言所开发的现代汉语自然口语语料库、自然对话语料库、现代汉语方言自然口语语料库; 中国科学院自动化所开发的旅游咨询口语对话语料库和旅馆预定口语对话语料库; 香港大学和香港理工大学联合开发的香港广州话语音资料库; 在汉语说话人识别领域应用较为广泛的863中文语音数据库[8]等等。

跨信道问题，一直是说话人识别领域一个非常重要的课题之一。NIST近年数据和LDC的许多专用数据库也都包括了不同场景、不同信道下的语音样本。浙江大学计算机学院录制的面向移动环境的SRMC[9]语音数据库和北京理工大学的BIT-Mobilespeech和BIT-MobileTalk[10]语音数据库增加了手机和pad的接入方式录音。但目前绝大部分语音库还停留在麦克风和固定电话两种录入方式。鉴于此，我们设计和建立了多路跨信道说话人识别语音库XMU-MultiChannel。

2 跨信道说话人识别语音库的建设

2.1 跨信道说话人识别语料设计

语音库的构建首先要从语料文本的设计入手，设计语料文本是一项高技巧的工作，包括语音库的完备性和自然性。语音库的完备性要求[11]是指，语音库要符合语言的概率模型，在保证文本真实性和口话自然度的前提下，用尽可能少的语句来覆盖所有的汉语发音现象。除了语音库的完备性，我们还要考虑语料的自然性。对话方式的语音是人类进行语言表述最自然的方式，它将包括篇章语音之外的一些语言和语音现象，如：情绪、心理变化对语音音调产生的影响，多发性的儿化音，口语化的语助词等。因此，我们按照汉语的口语习惯设计了跨信道说话人识别语料文本，包括：个人信息、数字串、短文和确定主题问答式对话四大部分。

个人信息包括姓名、性别、年龄、籍贯、学院、年级、学号、爱好共8项；数字串为录音人根据系统随机产生的3个数字串，每串用汉语读六遍，即18项，一个数字串由10 位数组成；短文部分共有20篇，每篇200字左右，每个录音人根据系统随机选的一篇短文，用汉语读一遍；确定主题问答式对话部分共有10个确定主题的话题，系统随机给出1个话题，要求一个录音工作人员根据话题对录音人提出多个问题，而录音人对每个问题都加以简要说明，采用一问一答的形式，时间至少为1 分钟。

2.2 采集系统

录音设备包括两台装有Skype软件和PowerGramo录音软件的电脑，其中一台电脑装有Steinberg Cubase 5录音软件、一个直式会议麦克风、一个头戴式耳麦、一个入耳式耳麦、一支录音笔、一部普通电话机、二部手机(一部为GSM网三星手机，一部为CDMA网中兴手机)、一个Modem调制解调器 (带有电话录音软件) 。该采集系统的采样率为44.1KHZ，精度为16bits，每个说话人采集的语音时长为3~4分钟。录音过程中，系统同时采集8种接入方式下的语音，第一路为会议话筒，第二路为入耳式耳麦，第三路为头戴式耳麦，第四路为CDMA网手机打入GSM网手机，第五路由网络即时语音沟通工具软件Skype拨打一个CDMA网手机号码，第六路由Skype拨打固定电话，第七路由录音笔录入，第八路由语音聊天软件YY语音录入，整个录音过程的接入和输出方式如图1所示：

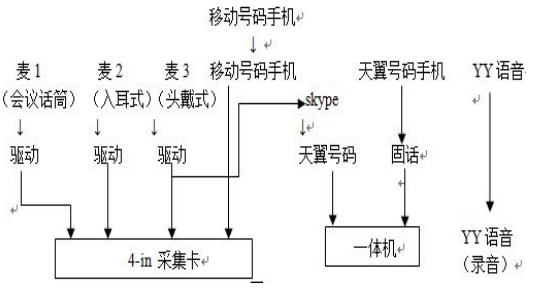

图 1 录音接入图

在进行语音信号处理之前，首先需要进行语音库的录制和建立。本文设计了一个语音录制系统，其整体的运行界面如下所示：

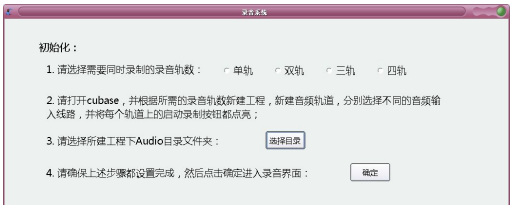

图 2 初始化界面

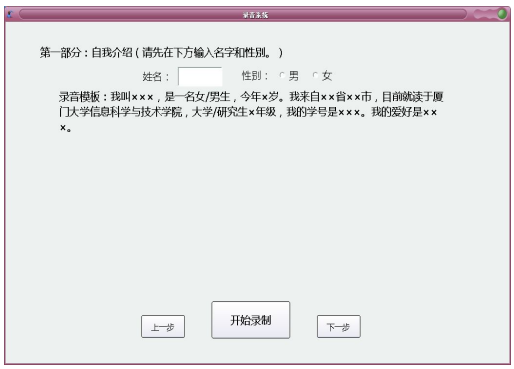

图 3 语音录制第一部分界面

图 4 语音录制第二部分界面

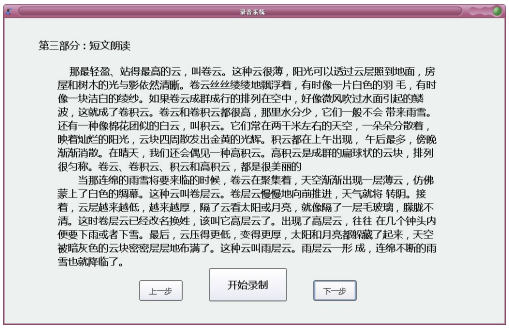

图 5 语音录制第三部分界面

2.3 录音过程

跨信道说话人识别语音库的说话人共有100名，为厦门大学在校本科生和研究生，来自全国各个省市，年龄介于19~ 25岁之间，男女比例为6:4。所有录音均在同一间安静的办公室(6m*5m)内进行的。录音要求为: 在自然放松的条件下，用正常语速和语调的普通话说出指定的语料。

录音前，先让录音人填写录音人登记表，录音人登记表包括录音时间、姓名、保存文件名、年龄、专业、籍贯、有无感冒/咽喉不适几项。然后再请说话人熟悉语料文本和录音系统。在各方面准备好后，进入录音系统开始录音，而旁边有专门的录音工作人员进行指导，并在录音的第四部分和录音人进行对话。

每次录音后，录音工作人员对语音文件进行人工检验，以排除录音过程中可能出现的错误。例如，查看语音文件的完整性并剔除语音中的信号过载音段、信号干扰、不规则噪声(例如咳嗽，机器噪声等)和非正常停顿造成的长时静音等。对于错误严重的录音文件，必要时可以请求说话人重新补录。

2.4 语音文件的命名和存储规则

录音数据是单声道的，其采样频率为44.1KHz，量化精度为16bits，保存格式为PCM方式的wav文件。每个说话人录的四个部分（自我介绍，数字部分，短文部分，对话部分）的语音直接保存在一个wav文件中，每个部分间由静音隔开。为了便于语音库的管理和数据共享，语音库的目录结构设计如图6所示。语音文件的命名规则是“说话人姓名_性别_信道编号_年_时间.wav”，共有8个信道，用“01”表示直式耳麦，“02”表示入耳式耳麦，“03”表示耳挂式耳麦，“04”表示移动手机打移动手机，“05”表示Skype打天翼手机，“06”表示天翼手机打固定电话，“07”表示录音笔，“08”表示Skype打Skype。例如，huangl_female_01_2012_152012.wav表示女性huangl在2012年15点20分12秒录的直式耳麦信道的语音文件。

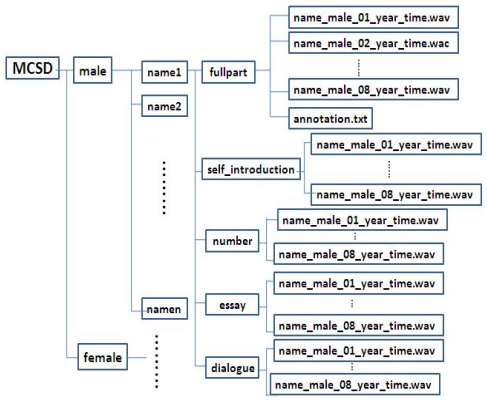

图 6 语音库的目录结构

2.5 语音后处理和标注

如何使跨信道说话人识别语音库发挥最大的作用，让所有的使用者能方便地共享数据，这是建立语音库的最终目的，也是我们必须考虑的问题。后处理的目的是将录制的原始语音库变成最终可以用来共享的语音库。后处理包括以下几个过程：

1）录音文件从录音设备向计算机传输。

2）语音数据的内容校对。

3）将不同格式的语音文件用CoolEdit软件统一转换成wav文件。

4）把同一说话人的8个不同信道语音文件用CoolEdit软件实现起止时间的同步。

5）语音标注

对语音进行标注是建立语音库的后期处理加工的一个重要内容，也是提高语音数据库质量和可用性的关键环节[1]。语音标注利用某种可书写文本符号来说明或描述语音波形中的语言现象，它体现了研究者对语音数据库信息的认知程度。每个说话人

都有一个.txt格式的标注文件，标注的信息包括以下几个方面：8个语音文件名各自代表的信道类型的标注；每个语音文件4个部分语音的起止位置的标定；语音段和非语音段的划分；各种突发噪声的标识。

3 基于 PLDA 的跨域迁移实验

目前，采用i-vector作为新的特征参数，使用PLDA建模可实现较高识别效果的跨信道说话人识别。但是，当开发集的语音样本过少时，就会严重影响PLDA模型的识别效果。假设已知有两个数据库A和B，其中A具有丰富的数据样本，可用于训练PLDA模型，而B则只有少量语音样本不足以独立训练获取对应的PLDA模型参数。若需要对数据库B进行测试，直接采用数据库A训练的PLDA模型参数，就存在数据不匹配问题。

为了解决这个问题，本文设计了一种实验方法，将原始数据库A的PLDA模型更新到目标数据库B中。

假设原始数据库A的PLDA模型参数为$\{\varphi_A,\Sigma_A\}$，目标数据库B的PLDA模型参数为$\{\varphi_B,\Sigma_B\}$，矩阵φ_A和φ_B描述数据库A和数据库B所表征的说话人相关子空间，Σ_A Σ_B为$\varepsilon_{i,j}$的方差矩阵，如公式（1）和（2）所示，采取线性加权更新的方法实现目标数据库B中PLDA模型的参数迁移：

$$\phi_r = \alpha\phi_A + (1-\alpha)\phi_B \tag{1}$$

$$\Sigma_r = \alpha\Sigma_A + (1-\alpha)\Sigma_B \tag{2}$$

φ_r为迁移后说话人相关的子空间，Σ_r为迁移后$\varepsilon_{i,j}$的方差，α为迁移调节参数。

4 实验结果与分析

本文使用厦门天聪智能软件有限公司[12]提供的实网数据做为原始数据库A，自建的跨信道说话人识别语音库作为目标数据库B。其中数据库A包含22000条语音，包含固定电话信道和麦克风信道。使用数据库A的2200条电话信道的语音（不分性别）训练512维的UBM，使用数据库A的12992条电话信道的语音（不分性别）训练一个400维的T矩阵，使用数据库A的6598条电话信道和麦克风信道的语音训练原始数据库A的PLDA模型ϕ_A和Σ_A；使用数据语音库B中的630条语音训练目标数据库B的PLDA模型ϕ_B和Σ_B。使用数据库B的60条语音做测试数据，数据库B的10条语音做target，共做了600次测试，其中，540次nontarget，60次target。再根据式(1)和式(2)，计算得到ϕ_r和Σ_r。使用这三个PLDA模型做测试，其DET(Detection Error Tradeoff))曲线比较如图7所示，其最小DCF数值如表1所示：

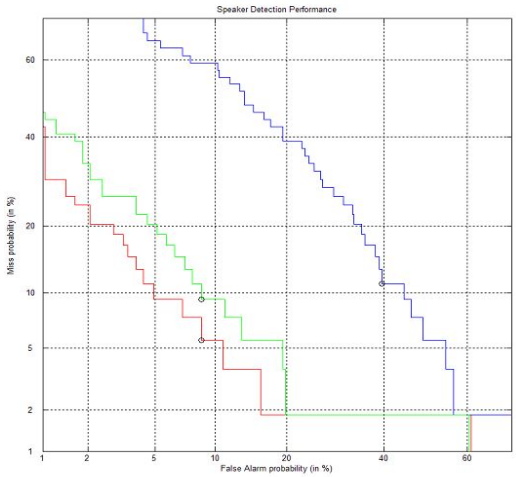

图7 DET曲线对比图

其中绿线为使用数据库B的PLDA模型φ_B和Σ_B做测试得到的DET曲线，其EER值为10%；蓝线为使用数据库A训练的PLDA模型φ_A和Σ_A做测试得到的DET曲线，其EER值为30%；红线为使用迁移得到的PLDA模型φ_r和Σ_r做测试得到的DET曲线，其EER值为7%。

表1 minDCF数值对比表

使用φ_A和Σ_A做测试	使用φ_B和Σ_B做测试	使用φ_r和Σ_r做测试
0.2675	0.0844	0.0610

通过比较图7中三条曲线的EER值和表1中的minDCF值，可以发现，使用迁移后的PLDA模型φ_r和Σ_r做测试可以获得更好的识别效果，使用φ_B和Σ_B做测试获得的识别效果不是很理想，是由于目标数据库B数据有限，训练得到的PLDA的模型不能很好的表征说话人的语音特征；而使用φ_A和Σ_A做测试效果最差，是由于训练PLDA的数据库和做测试的数据库数据不匹配。而φ_r和Σ_r是通过迁移得到的，融合了φ_A、Σ_A和φ_B、Σ_B各自的优点，具有更好的识别效果。

5 总结

本文首先介绍了跨信道说话人识别语音库XMU-MultiChannel的建设，这个语音库含有100个说话人，每个说话人包含8个信道的语音，含有丰富的信道信息。为验证语音库XMU-MultiChannel的有效性，本文实现了基于PLDA跨域迁移的说话人识别实验，使用大样本数据库训练的PLDA模型参数与小样本数据库XMU-MultiChannel训练的PLDA模型参数，融合得到新的PLDA模型参数，以提高基于跨信道小样本数据库的说话人识别性能。

参 考 文 献

[1] 王蕴红. 863 申请项目报告[R]. 2001.

[2] Joseph P. Campbell. Jr. and Douglas A. Reynolds. Corpora for the evalution of speaker recognition systems[C]. Acoustics, Speech and Signal Processing, 1999, 2:829-832.

[3] Hakan Melin. Databases For Speaker Recognition[C]. Activities In Cost250 Working Group 2, 1999.

[4] Cole R. Noel M, Noel V. The CSLU speaker recognition corpus[C], Acoustics, Speech and Signal Processing, 1998, 3167-3170.

[5] Falcone M, Gallo. The SIVA speech database for speaker verification: description and evaluation[C], Acoustics, Speech and Signal Processing, 1996, 1902-1905.

[6] J. Hennebert, H. Melin, D. Petrovska, etal. POLYCOST: A telephone-speech database for speaker recognition[J]. Speech Communication 2000, 31:265-270.

[7] Javier Ortega-Garcia, Joaquin Gonzalez-Rodriguez, Victoria Marrero-Aguiar, AHUMADA: A large speech corpus in Spanish for speaker characterization and identification[J], Speech Communication, 2000, 31:255-264.

[8] 周昊郎, 王岚, 吴玺宏, 迟惠生. 一个面向说话人识别的汉语语音数据库. http://nlprweb.ia.ac.cn/english/irds/chinese/SinobiometricsPDF/Wuxihong.pdf,2002.

[9] 杨莹春, 颜时锋, 吴朝晖, 桑立锋. 面向移动互联环境的说话人识别语音库 SRMC[C]. 第七届全国人机语音通讯学术会议, 2003, 238-242.

[10] 尹安容, 黄石磊, 李代松, 陈嘉, 谢湘. 基于电信网络说话人识别语音库的设计和实现[C]. 第八届全国人机语音通讯学术会议,2005, 495-497.

[11] 王侠, 吴及, 肖熙, 王作英. 关于语音库建设若干问题的讨论[C]. 第六届全国人语音通讯学术会议,2001.325-348.

[12] http://www.talentedsoft.com/.

The Design and Research of Cross-channel Speaker Recognition Database

Lin Li, Lihong Wan, Ling Huang, Qingyang Hong

School of Information Science and Technology, Xiamen University, Xiamen 361005, China

qyhong@xmu.edu.cn

Abstract: The establishment of a Chinese speech database has a great significance in promoting the research and application of speaker recognition technology. This paper designs a cross-channel speech database XMU-MultiChannel for the research of channel mismatch in speaker recognition. After comparison of existing domestic and international speaker recognition speech database, this paper has a discussion on recording program, acquisition system, recording content, labeling and storage. Based on our designed XMU-MultiChannel database, a cross-domain transfer strategy experiment of PLDA was implemented. As a result, the difficulties for small samples to train the PLDA model could be resolved.

Key words: speech database; speaker recognition; cross-channel; cross-domain transfer