维吾尔语口语语音语料库的设计与研究

杨雅婷[1,2]，马博[1,2]，王磊[1,2]，吐尔洪·吾司曼[1]，李晓[1]

1. 中国科学院新疆理化技术研究所，乌鲁木齐 830011；
2. 中国科学院研究生院，北京 100190
E-mail: yangyt_xj@sina.com，Phn: +86-15026043990

摘 要：在分析维吾尔语语音语料特点的基础上，结合实际语料库建设需求和地域语言特色，提出了适用于维吾尔语口语语音语料库建设的语料库设计规范、语料内容、语音采集和标注方法，并就不同信道对语音特征参数的影响进行分析。研究拟建立时长 300 小时的维吾尔语口语语音语料库，有效改进少数民族语言语音学的资源研究和基础建设。

关键词：维吾尔语；口语；语料库；特征参数；语音识别

Research on the Uyghur Spoken Language Speech Corpus

YANG Ya-ting[1,2]，MA Bo[1,2]，WANG Lei[1]，TURGHUN Osman[1]，LI Xiao[1]

1. Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011;
2. Graduate University of Chinese Academy of Science, Beijing 100190
E-mail: yangyt_xj@sina.com，Phn: +86-15026043990

Abstract: Based on analyzing the general construction method of speech database, combined the actual requirement and region language characteristics, provided the design criterion, database contents, collect and mark methods to establish the Uyghur spoken language speech corpus that include 300 hours speech data. Then, present a research on the effect of the channel on the parameters of speech signal.

Key words: Uyghur; Spoken language; Corpus; Speech parameter; speech recognition

1 引言

口语语料库的建设和积累是进行计算语言学研究的基础资源。根据语料载体划分，口语语料库又可以分为“文字模式”和“语音、文字共存模式”两种，本文针对后者进行研究。近年来，国外发达国家十分重视语音语料库的建设，欧洲共同体 90 年代初建立了基于欧洲七种语言的语音语料库研究计划“EUR-ACCOR”，关注多样化语音资源的建设。国内的中科院声学所、清华大学、社科院语言所等研究机构，对其也进行了广泛研究。在国家自然科学基金、社会科学基金和各部委研究基金的支持下，少数民族语言语音语料库正在建设中[1]。

新疆地区是多民族聚居地，其中维吾尔族人口数约占地区总人口数的 46.1%，拥有丰富的少数民族语言资源。基于地域和语言特色，从灵活性和实用性出发，展开维吾尔语（后文简称维语）语言学的基础研究，建立维语口语语料库，填补了我国少数民族语言语音语料库中的部分空白。本研究将有助于提高少数民族语言语音基础研究和应用研究水平，保证科研资源的共享和科学研究的延续性，能够有效改进我区少数民族语言语音学研究方法和手段，

基金项目：中国科学院“西部行动计划高新技术项目”（The western high technique program of Chinese Academy of Science. No. KGCX2-YW-507）；中国科学院“西部之光”项目。

作者简介：杨雅婷（1985-），女，博士生，主要研究方向：多语种信息处理技术；马博，博士生；王磊，博士生；吐尔洪.吾司曼，助理研究员，硕士；李晓，研究员，博导。

既加速了推进弱势语言的标准化、规范化、信息化和现代化进程，同时还维护了“语言的多样性”。

2 语料库的特点

目前，维语语音语料库的建设尚处于起步阶段，已建设的语料库主要是维语标准发音库。这些语音语料库多在环境良好的实验室中录制，以固定内容的朗读、新闻播音为主，对实验语音学、语音识别的基础性研究起到了重要的作用。但其灵活性、工程实用性较差，用于自然口语的研究时，其识别率较低。语料库类型及其实用性如图 1 所示。因此，设计和采集维语口语语料库能够为维语的语音信息化建设、语音教学、语音通讯、语音识别、语音合成等自然对话系统提供真实有效依据[2]。

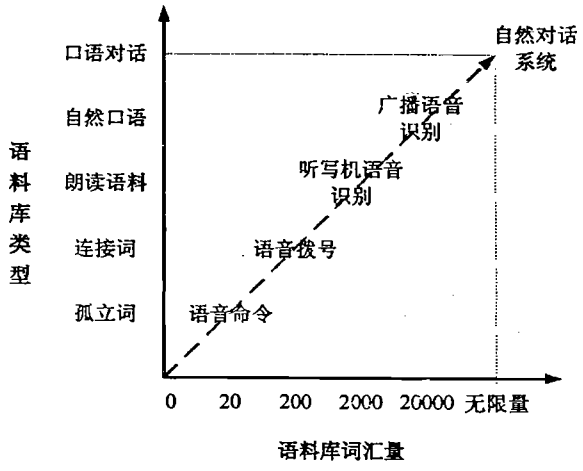

图 1 语料库类型及其实用性

2.1 维语语音语料的特点

维语属于阿尔泰语系突厥语族西匈语支，在语法上属于黏着语类型。它的音素、音节等发音单元具有本质发音特点。维语语音有元音 8 个、辅音 24 个。由辅音和元音构成维语语音音节，每个音节必须且只能有一个元音[3]。维语音节的三大块是:(起音)+领音+(收音)。如果用字母“V”代表元音，“C”代表辅音，维语的音节可以归纳为以下几种形式：V、VC、CV、VCC、CVC、CVCC。部分音节在语流中产生语流音变现象，常见的有同化、弱化、脱落以及元音和谐等现象。维语的发音规律和语音现象有很鲜明的特点，其元音、辅音以及语音结构的最小单位音节的识别对维语语音识别有很重要的意义。

维语中约有 5000 多个音节，包括外来语借词。所谓外来语借词是指在民族的交往接触中，遇到本民族没有的事物和概念时，为了称呼它们，可以利用本民族语言的构词材料构成新词，也可以直接吸收外来词语，那么这种从外来语中吸收的词就叫借词，也叫做外来词。它们的出现对语料标注提出了新的规范。

2.2 口语语料的特点

语音语料库的类型从话语的自然程度可分为朗读语音语料库和自然口语语音语料库。朗

读语音语料指有计划的、按文字朗读的语音库；自然口语语音语料库指没有计划的自然口语对话或独白语篇语音库。随着口语人机交互系统技术的不断发展，口语语料在系统设计中变得越来越重要。一个语言片段就是一段真实生活的再现，而实现该目标的途径之一就是建立音文同步的口语语料库。

口语是人们在进行自然交流时的语言现象，口语语料库的数据要求能够充分反映口语现象，有助于进行自然对话语音识别。自然口语语料录制双方的自由谈话内容，不指定文本，该数据自然度很好，能够包含比固定语料更多的语言和语音现象，如：情绪、心理变化对语音音调产生的影响，多发性的儿话音、口语化的助词等。但是，口语语料库数据量大，语料的收集、整理、标注和维护工作量也很大，管理较为困难。

3 语音语料库的设计

鉴于该语料库的特性,将其设计分为：语料库设计规范、语料内容设计、语料采集、语料存储、语料标注和特征参数分析六个阶段，依次进行研究[4]。维语口语语料库建设过程如图 2 所示。

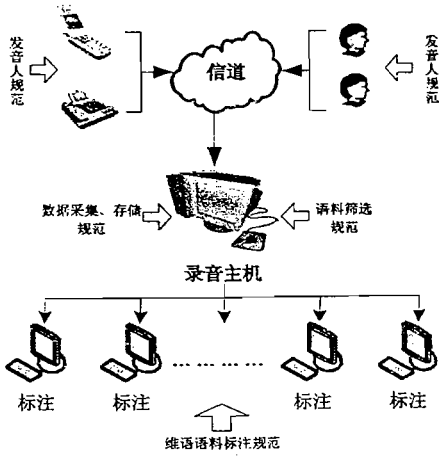

图 2 维语口语语料库建设过程

3.1 语料库设计规范

本研究致力于建立维语统一的、完备的口语语料库。为达到数据库冗余度小和代表性、实用性强等特点，语料库设计之初需制定相应的发音人、语料筛选、数据录制、数据存储、语料标注规范。构建该语料库不仅要考虑发音人的分布、数据规模的设计[5]。语料库设计规范如表 1 所示。

表 1 语料库的一般规范

规 范	内容说明	具 体 规 范
发音人规范	发音人信息、要求	年龄、性别、方言背景及发音注意事项。
语料筛选规范	语料的组织、规模	电话录音的长度及用词要求。
数据采集规范	录音设备、环境指标	使用声音采集软硬件电话信道采集。
数据存储规范	采样率、存储规范	8KHz 采样率，.wav 格式存储。
语料标注规范	标注内容、说明	使用规范字体对说话人、语音内容进行标注。

3.2 语料内容设计

建立一个口语语料库的目的是为了更好地进行口语语音识别，因此，口语语料库中的数据应该能够充分反映口语现象，尽量覆盖连续语流中，受到韵律因素的影响，音节音段音联表现的复杂性。而在实际语料收集时，人们总是自觉不自觉地改变了讲话的习惯，使得收集到的语料不再能够充分反映口语现象。为了能收集到既能反映人们口语现象又能限制在一定任务领域内的实际语料，可以包含对话、独白两部分语料。

（1）对话（人人对话）

对话双方就某一个或几个人、物、问题进行交谈和讨论，录制双方交谈过程的所有语音。对话过程中，语流音变现象出现机率较高，且融合情感变换、语调变换，适用于语音识别中的情感识别。语料内容贴近生活，体现出人与人交互的语言习惯，融自然与实用为一体，适用于自然口语语音识别。

（2）独白（人机对话）

由于心理预期的作用，人机讲话时的语言习惯与人人交互不一样。因此在收集口语语料时，不仅要收集人人对话时的语料，而且还要收集人机对话的口语语料。根据一个给定的主题或一个固定的提问，进行口语化的独白阐述，其设计目标是包含丰富的维语语言特征。针对中国少数民族语中，使用外来语借词的共性，特定的提问。独白部分所采集到的语音将属于主题相同的语料。包含特定领域的内容，为面向领域的语音识别提供语料。

3.3 语料采集

本研究采集方式可以分为麦克风信道和电话信道，以 8KHz 采样率直接进行语音采样，并以.wave 格式存储。麦克信道用 Cooledit 录制，电话信道采用声音采集软硬件同时对 2 路电话录音，主控计算机将显示整个录音过程，录音管理者可通过同步监控发现录音过程中的问题，并及时处理。拟采用系统工作流程图如下图 2 所示。

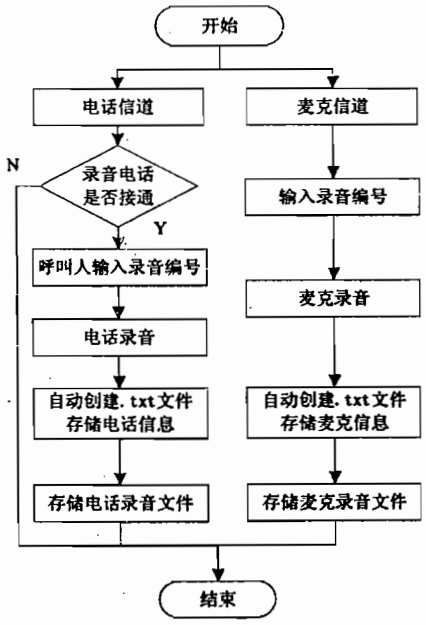

图 2 系统工作流程图

这两种方式采集到的语音质量是不一样的。麦克风方式采集的语音质量更好，因为原始语音未经任何处理即被采入，频谱上没有什么变化。而电话采集由于线路噪声以及电话线路传输带宽的原因，波形上会发生一些变化，尤其是通过手机，由于传输时经过压缩，现象更明显一些。但是，在麦克风方式下由于话筒和说话者的距离发生变化的可能性较大，音场会发生变化，声音随着说话人的移动有着强弱变化，且环境噪声有时也会很大，在电话方式下这些问题基本上不会存在。

3.4 语料存储

每个语音语料文件至少存贮在不同的两种存储介质上，并对数据建立存储信息：一是发音人属性，如发音人年龄、性别、方言背景等；二是实际语音语料，主要用于保存录制好的语音波形图形的原始参数；三是发音文本，标注实际语音语料对应的文本[6]。

（1）每位发音人对应一个描述文件，记录发音人的信息：

Speaker ID Sex Age Dialect Recording date

（2）每一个声音样本对应一个描述文件：

Record ID Speaker ID Recording date Recording place Environmental Conditions Channel Sampling rate Bits per sample Corresponding annotation file

3.5 语料标注

语音语料库不仅记录语谱图等语音学数据，还有各种语言学信息标记[7]。适当规模的语料库经过科学选材和标注，能够反映和记录该语种语言的实际使用情况。本研究首先对现有原始语料库，根据需求进行具体设计，采用 Praat 专业软件采集进行规范标注。主要的标注内容为 SPEAKER 层标注、CONTENT 层标注 ，如图 3 所示。

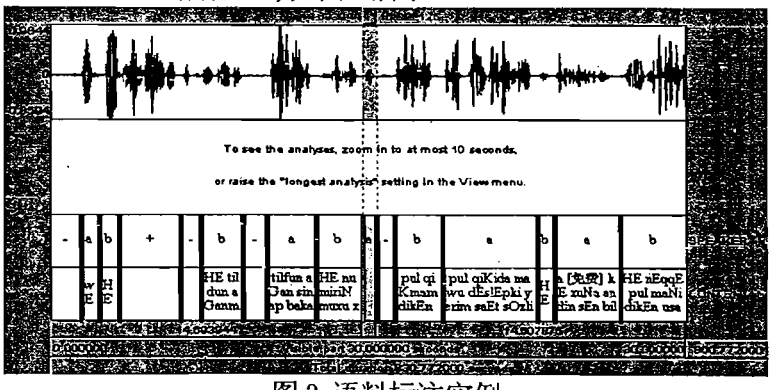

图 3 语料标注实例

1. SPEAKER 层上的合法标注

（1）a，b，c …：表示不同的说话人；

（2）+：表示两个或多个说话人同时说话，发音重叠；

（3）-：表示无效语音，例如听不清的语音，咳嗽、笑声、噪音等。

2. CONTENT 层上的合法标注

（1）如果单词的发音有脱落，用用圆括号把脱落的部分括起来，例如：شۇڭا (خا)زىر；

（2）词与词之间用空格隔开；

（3）数字也要标注为维语字母，不能使用阿拉伯数字，例如：بەش（阿拉伯数字 5）；

（4）SPEAKER 层标记为“-”或“+”的部分，CONTENT 层不进行任何标记；

（5）外来语借词用方括号括起来，例如：[MP3]。

（6）拟采用语音标注规范字体如表 2 所示。

表 2 拟采用语音标注规范字体

维文字母	拉丁字母	维文字母	拉丁字母	维文字母	拉丁字母	维文字母	拉丁字母
ئا	a	ب	b	ھ	H	ق	N
ئە	E	پ	p	خ	h	گ	g
ئې	e	ت	t	ل	l	ك	k
ئى	i	ن	n	م	m	ز	z
ئو	o	ج	j	ق	K	ي	y
ئۇ	u	ر	r	ە	d	ف	f
ئۆ	O	س	s	ۋ	w	غ	G
ئۈ	v	ش	x	چ	q	ژ	Z

3.6 特征参数分析

语音特征参数是语音识别、说话人识别研究的基础，对识别率有着直接的影响。本研究采用了麦克信道和电话信道采集口语语音语料，不同的信道对语音信号产生影响也不同，语音参数也会发生相应的变化。通过电话信道之后语音线性预测系数或其派生参数，频率直接导出参数，混合参数，都会不同程度的收到一定的影响。电话信道对极点的影响严重，语音信号经过电话信道的干扰，会产生失真现象。声音产生过程中，声道的全极点模型可以表示为：

$$H(z)=\frac{1}{A(z)}=\frac{1}{1-\sum_{i=1}^{p}a_i z^{-1}}=\sum_{i=1}^{p}\frac{r_i}{1-f_i z^{-1}} \tag{1}$$

其中，$A(z)$ 为线性预测多项式，f_i 为第 i 个极点；r_i 是第 i 个极点的留数。每个极点 f_i 与其中心频率 ω_i 和带宽 B_i 相关：

$$f_i=\exp(-(B_i-j\omega_i)) \tag{2}$$

这样，每个极点都可以由参数（ω_i，B_i，r_i）表示出来，共振峰通常由带宽较窄的极点成份表征出来，而带宽较宽的极点则反应了信道和声门特性。电话信道对语音信号频谱包络的影响如图 4 所示。

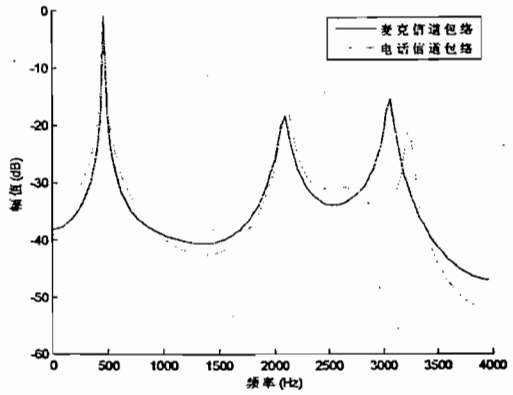

图 4 电话信道对语音信号频谱包络的影响

可以看出，在经过电话信道后，语音信号的频谱发生变化，具体表现在共振峰位置和宽度会发生变化，反映出电话信道的线性影响；除此之外，经过电话信道的语音频谱包络会出现虚峰现象，说明了电话信道对语音信号非线性影响的存在。

因此，本研究在进行特征参数分析时将线性预测倒谱系数(LPCC)、梅尔倒谱系数(MFCC)、自适应成分加权特征(ACW)倒谱等语音特征参数进行对比[8]，并考虑了语音信号的动态特征(瞬变特征)，形成了动态信息和静态信息互补。本研究中语音特征参数分析具体方法已经另文详细阐述，不再赘言。

4 结束语

本研究拟建立包含时长 300 小时语料的维吾尔语口语语音语料库，该语料库的建立能够有效改进少数民族语言语音学的资源研究和基础建设，为计算语言学的基础研究奠定基础，并对少数民族语言语音特征参数进行分析研究及语音识别技术研究起到了促进作用。

参考文献

[1]Huhe,Haschimeg,Zhou Xuewen,et al.Symposium on National Minority Speech and Language Processing Technology: A develop method of chinese national minority speech and language parameter database [C] //Proc of NCMMSC'2009,Lan Zhou,2009:555-560.

[2] Rabineer L R，Juang B H. Fundamentals of Speech Processing and Recognition[M].Prentice-Hall,1993.

[3]Gulijiamali Maimaitiaili，Aisikaer Aimudula. The Phoneme Feature Based Uyghur Speech Synthesis[J]. Journal of Chinese Information Processing, 2008,22(4):100-104.

[4] Xu Yonghua,Yang Jian,Chen Jiang,et al.Symposium on National Minority Speech and Language Processing Technology: A telephone speech database of national minority speech and language [C] //Proc of NCMMSC'2009,Lan Zhou,2009:410-413.

[5] Askar Hamdulla,Dilmurat Tursun.An acoustic Parametric Database for Uyghur Language[C] //Proc of 2009 International Joint Conference on Artificial Intelligence,IEEE,2009:405-408.

[6] Hennansky H. RASTA Processing of Speech[J]. IEEE Transactions on Speech and Audio Processing，1994，2(4)：578-589.

[7] 蔡莲红，黄德智，蔡锐等.现代语音技术基础与应用[M]. 北京：清华大学出版社，2003.

[8] Assaleh K T，Mammone R J. New LP-derived Features for Speaker Identification[J].IEEE Transactions on Speech and Audio Processing,1994,2(4):630-638.