维哈柯及蒙语多文种语言相似性考查研究

达瓦·伊德木草[1,2] 艾尼宛尔·托乎提[2*], 于清[1*]， 吾守尔·斯拉木[1,2]

[1]新疆大学信息与工程学院， 乌鲁木齐， 中国
[2]新疆多语种信息技术实验室，乌鲁木齐， 中国

摘要：本文以阿勒泰语系下的维哈柯及蒙古语多语言平行文本和语音语料为研究对象，分别对比多语言文本量化序列向量及语音声学音律特征的相似度，研究语言信息间存在的相通性。试验发现，同语系同语族黏着语言相似度较高：文本相似性达 85%；声频特征相似性达 95%。从而确认在同语系多种黏着语言间创建语言信息共享云模的可行性，这将有利于实现语言文本及语音信息的跨语言转换处理，极大降低少数民族语言信息处理成本。

关键词：同语系同语族语言；平行语料；声学音律特征；基频 F_0；相似性考查

An investigation research on the similarity of Uyghur Kazakh Kyrgyz and Mongolian languages

[1]College of information science & engineering Xinjiang University, Urumqi, China

[2]Xinjiang Laboratory of Multi-language Information Technology, Urumqi, China

Abstract: In this paper, an investigation is done for the similarity between the same family and agglutinative languages (such as Altai family languages ,for example, Uyghur, Kazakh, Kyrgyz and Mongolian using different countries and areas). Cosine similarity measure is used to calculate the similarity using the parallel texts and the acoustic features extracted from the same content speech sentences spoken by the different language speakers. Experimental results show that the transformation is more feasible by word to word units when learning the connection rule of a stem and an affix (function words) between languages by word level and common acoustic models. Thus, this avoids the uphill work of MT for the resource-deficient languages such as minority languages being used in the developing countries. Additionally, the costs can be reduced.

Keywords: same family and agglutinative language, parallel text, acoustic and prosody parameters, $F_{0,}$ similarity.

1. 引言

多语言信息处理，尤其是少数民族语言信息处理正从文字信息处理阶段跨越到较复杂的自然语言及语音处理阶段，机器翻译 MT(Machine Translation), 大词汇连续语音识别 LVCSR（Large Vocabulary continuous speech recognition ）等新技术在少数民族语言信息处理中逐步得到预期测试效果[1,2,3]。

语言信息的自动处理往往需要丰富的语言信息知识，大规模语言资源的收集、整理、建设，需要耗费大量人力、物力、财力，并且对于小语种语言（即少数民族语言）其现有语言资源

This research is sponsored by NSFCP61163030. *Corresponding authors: Aniwar and Yuqing e-mail: misiran@foxmail.com, yuqing0131@126.com

缺乏，严重阻碍了少数民族语言信息处理的深入发展。本文研究同语系多种黏着语言间的相似性，以期实现语言资源间的共享。

自然界存在许多较相似的言语，如同语系语言，而同一语系下同语族语言间相似性更高，这些语言不仅在文字字模，构词方法，语序，句法，语法等结构上较接近，而且在发音风格上有更多相似特征[4]。接下来将以阿勒泰语系下土耳其语族 TLB（Turkish Language branch）和蒙古语族 MLB（Mongolian language branch）的文本信息为例进行说明。图-1 显示了维（Uyghur）哈（Kazakh）柯（Kyrgyz）三种语言的文本句对，及其相应的 Unicode 编码，三条语句都表达“你什么时候来我们家？”，它们同属土耳其语族。仔细观察发现，每条语句由若干个阿拉伯字母按至右向左顺序书写而成，字符串间用空格分隔。虽然有 Uyghur, Kazakh, Kyrgyz 不同语言之分，但其字模，字符串构成方式，语序以及句法和语法规则大体相通。另外，三种语言对应字符串的 Unicode 编码不仅内容上大体相同，而且在表现形式上(斜体字部分) 也较接近，即使某些略有差别，但切分词干与后缀功能词后，词干部分几乎相同。如图-1 各条语句的第一个字符串(从右)编码中，词干/biz/都相同，仅后缀功能词不同。黏着语言中这些功能词数量有限，这充分说明同语族语言在书写表现形式上有公共信息。

图-1　维哈柯语文本句对及其 Unicode 编码

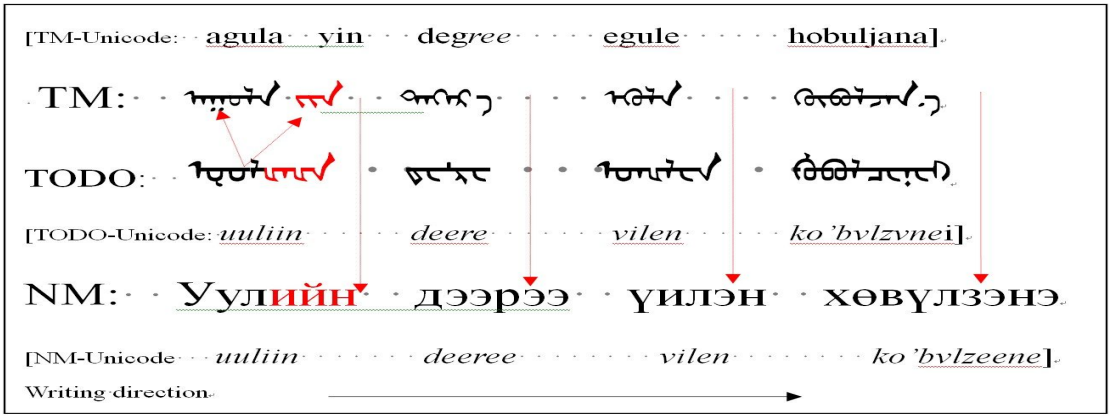

图-2　不同蒙古语文本句对样式及其 Unicode

这种公共信息结构也同现于蒙古语族，图 2 显示了三种蒙古语(TM，TODO，NM)文本句对样式，它们同属 MLB，目前在不同国家或不同地区被使用。观察它们的 Unicode 编码，发现 TODO 与 NM(New Monglian 蒙古国语言文字系统)语言词对齐公共部分出现较多。图-3 进一步说明 TODO 与ＮＭ词与词之间直接转写的可能性较大。

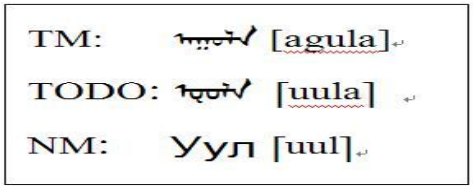

图-3. MLB 语言间词对齐关系

据以上分析，同语族各语言间存在较多公共信息，能否有效利用这些公共部分实现各语言之间的文本语音信息的转换处理，从而降低少数民族语言与不同语序，不同语法语言(如汉语)之间的翻译处理难度，是极其有意义的讨论课题。据此本文设计以下技术路线，如图-4 所示，先采用 MT(Machine Translation) 高代价复杂技术解决汉语与维语的转换问题，再讨论用 TT(Text Transformation)技术解决同语族语言文本转换问题，进而实现汉语与不同少数民族语言的机器翻译。该方法或许比各少数民族语言单独使用 MT 技术更方便有效。为此，探讨语言之间共享性或者相通性很有必要。

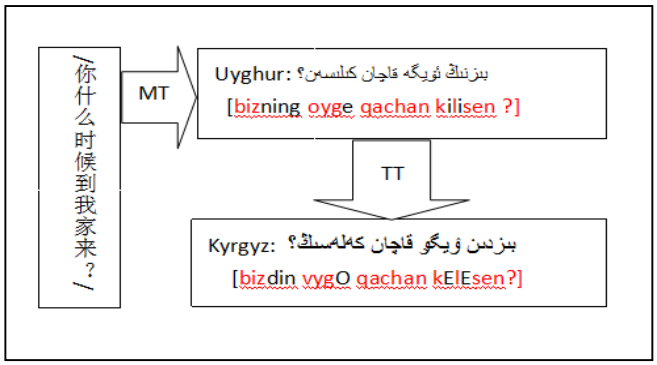

图-4. 汉语与少数民族语言（同语族语言）机器翻译技术路线

本文组织结构如下：第 2 节简介相关研究现状，第 3 节讨论 Cosine 相似尺度理论，第 4 节基于对齐文本及语音音律参数，利用 Cosine 相似度算法，通过具体实验考察各语言间相似性，分析实验结果，第 5 节结论与展望。

2. 相关研究现状

近年来，关于跨语言信息处理研究，主要侧重于跨语言检索以及相似语音参数横向移植等方面。文献[5]运用德英法等 15 种欧洲语言语音声学参数横向移植，实现目标语的语音识别。文献[6]借助机器翻译实现中文与英文文本跨语言信息检索。文献[7,8]阐述了在同一语言文本中，通过计算句子相似度，获取语义接近的句子，提高机器翻译质量的方法。然而，关于相似语种的文本及语音信息的横向转换处理研究，还很稀少。本研究前期工作基于语料库以及语言学规则实现蒙古语多文种横向转写，取得较好成果[9,10]。

3. cosine 相似尺度

设有两个 n 维向量 A 和 B，如公式(1)所示，这两个向量的相似性由公式(2)给出。当 $\cos ine\theta = 1, (\theta = 0^0)$ 时，两个向量 A 和 B 相同，即 A 和 B 完全相似；当 $\cos ine\theta = 0, (\theta = 90^o)$ 时，两个向量 A 和 B 完全不相同，即 A 和 B 无相关性；用 $\cos ine\theta$ 在[0,1]之间的取值，度量两个向量 A 和 B 的相关程度[11]。

$$\begin{cases} A = a_1, a_2, \cdots a_n \\ B = b_1, b_2, \cdots b_m \end{cases} \quad (1)$$

$$sim = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{m} (B_i)^2}} \quad (2)$$

4. 相似度考察实验

4.1 文本相似度考察实验

4.1.1 实验数据

本实验的数据来源于多语言平行文本语料，该语料由科研项目 NSFCP61163030*支持建造，有关该语料的数据统计信息见表 1。

表 1 多语言平行文本语料数据统计信息

Language（abbr.）	#of sentence	#of entry	#of stem	#of alphabet
Chinese（cn）	7854	101,235		
Uyghur（u）	7654	45,872	8626	32
Kazakh（h）	7854	38,050	7508	27
Kyrgyz（ky）	7854	34,545	8658	32
TM（tm）	7854	31015	12833	31
TODO（todo）	7854	32420	9531	31
NM（NM）	7854	28043	8416	34

4.1.2 实验方法

首先对语料中各种语言的文本句对进行量化处理，获取量化向量归正参数；再利用公式(2)分别计算句对级以及词对级相似度。

4.1.3 实验结果及分析

图 5 显示了各语言句对级相似度计算结果，从图中观察到，在文本级实验中，同语族语言之间相似度较高，MLB 语言之间相似度达到 0.8，TLB 语言之间相似度高达 0.9；不同语族的语言之间相似度明显下降，如 TLB –TODO，TLB –TM；并且 TLB –TODO 语言(同地区不同语族语言)的相似性略高于 TLB –TM（不同地区不同语族语言）。

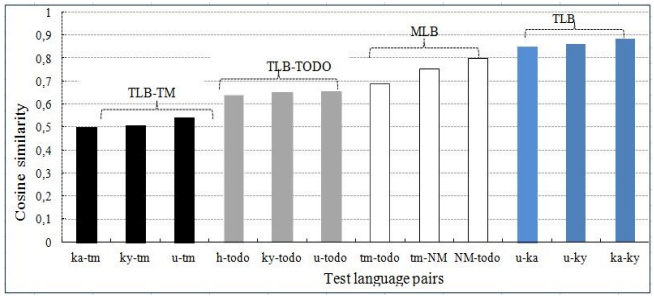

图-5 各语言句对级相似度计算结果

上述同语族语言之间以及不同语族语言之间的文本相似度差别，同现于各语言词对级相似度计算结果中，并且表现得更加明显，如图-6 所示。图中显示 MLB 词相似度接近 0.9，TLB 词相似度超过 0.9，然而不同语族语言之间词相似度极低。实验结果揭示，对于不同的少数

* National science fund of China and Science fund of Xinjiang government

民族语言，如果它们属于同一语族，则实现不同形式语言文本转换处理，在词级单元平行进行是可能实现的。

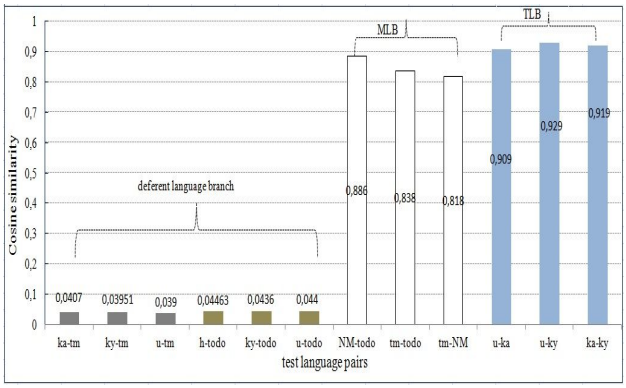

图-6 各语言词对级相似度计算结果

4.2 语言的发音相似度考察实验

4.2.1 实验数据

本实验以维哈柯语言为主，利用平行语料录制语音，分别选用各语言 10 个发话人，每人朗读相同内容的 50 个句子，进行录制。录制数据用 16KHz，16bit，单声道 WAVE 格式保存。最后，对录制的每句语音流，人工严密地标注出音素，再分别抽出音素单元的声学特征参数以及句子发话段的基频参数 F_0，见图-7，本实验将分别考察各语言声学特征及音律特征的相关性，进而探讨相似语言语音信息横向处理的可行性，这将有利于相似语言连续语音识别，语音合成等跨语言信息处理的深入发展。

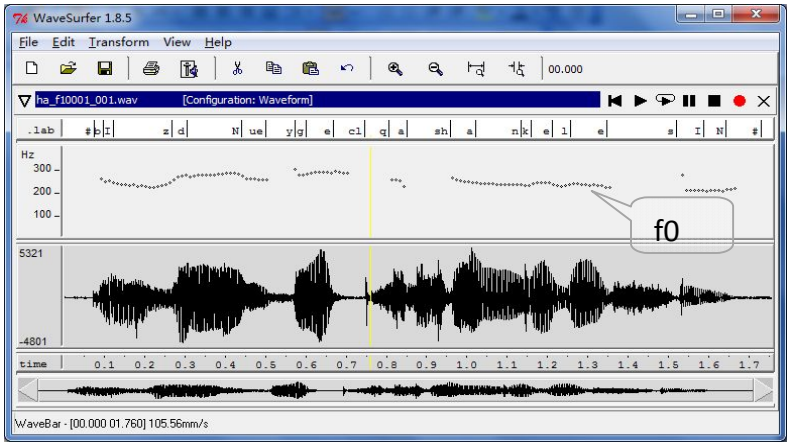

图-7 声频分析

4.2.2 共振峰分析

共振峰是指说话者声道脉冲响应，如果将声道视为一个谐振腔，共振峰就是这个腔体的谐振频率。表示浊音信号，最主要的是前三个共振峰 F_1,F_2 和 F_3(见图-8)。本实验利用 LPC(频域线性预测算法)，提取元音前两个共振峰 F_1 和 F_2，分别比较 TLB 语言和 MLB 语言的声频特性。TLB 语言和 MLB 语言基本元音的 F_1 和 F_2 共振峰分析结果分别见图-9(a,b,c) 和图-10(a,b)，为比较黏着语言常用标准，图-9(d)中给出日本语五个元音共振峰标准分布图[13]。分析以下各图，得出结论：① 同语族语言 TLB 中各元音 F_1 共振峰取值大致相同(350Hz~950Hz)，F_2 共振峰有明显差别，哈语和柯语取值范围明显高于维语，维语为 500Hz~4000Hz，而哈语为 900Hz~5000Hz，柯语为 100Hz~7000Hz。

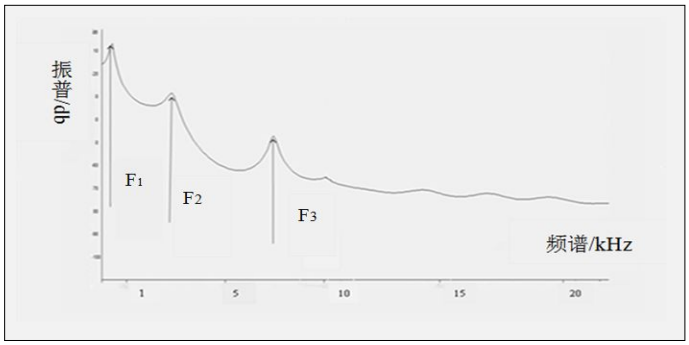

图-8 元音共振峰提取方法

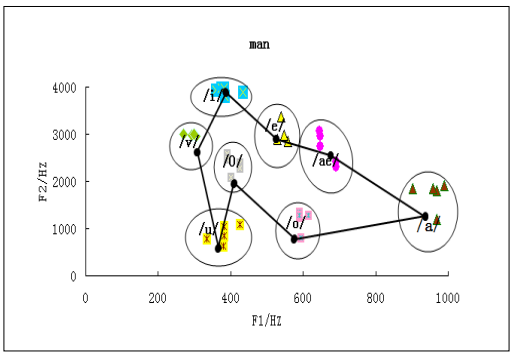

图-9(a)维语男声八个元音 F_1 和 F_2 分布

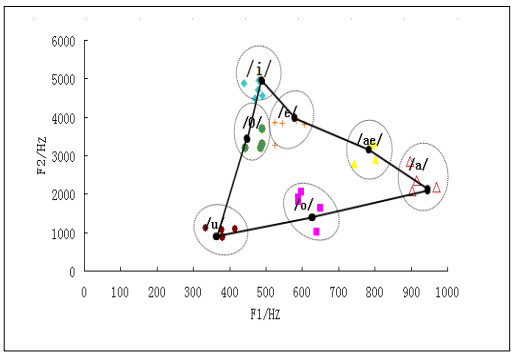

图-9(b)哈语八个元音 F_1 和 F_2 分布

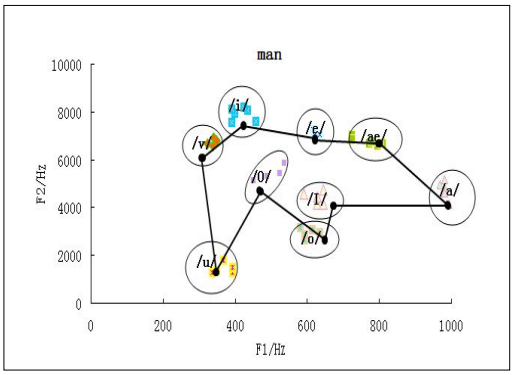

图-9(c)柯语九个元音 F_1 和 F_2 分布

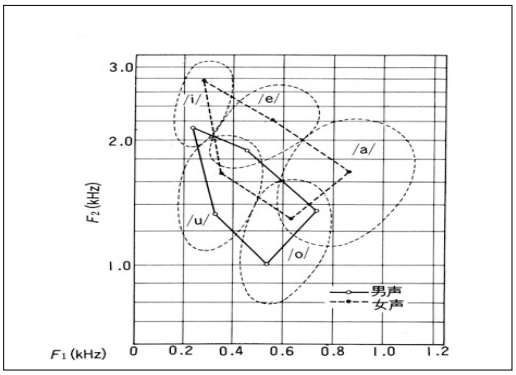

图-9(d)日本语五个元音 F_1 和 F_2 分布

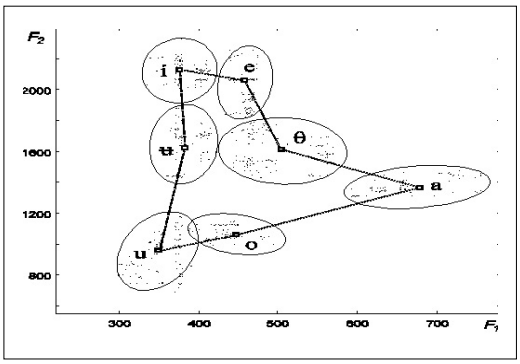

图-10(a)蒙语新疆地区发音七个元音 F_1 和 F_2 分布

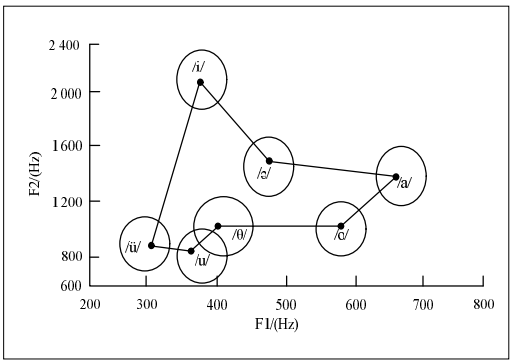

图-10(b)蒙语内蒙地区发音七个元音 F_1 和 F_2 分布

②比较图-9 和先行研究结果图-10[14]，不同语族(TLB 和 MLB)语言的基本元音共振峰分布特性差别较大，并且从图-10（新疆和内蒙地区蒙语口语发音）观察到不同地区的蒙古语发

音有明显差距。

4.2.3　　音律特性—基频(F_0)分析

人类的语音信息主要体现在韵律的变化上，在韵律特征中，基频结构最能反映说话人的语言信息特征。语音中只有浊音和元音有周期性脉冲串，其频率就是基音频率，简称基频F_0。实验利用语音信号时域算法工具 Wavesurfer 提取不同语言发话段的基频 F_0 曲线，分析比较各语言基频之间的相似性。表 2 和表 3 以及图-11(a)和图-11(b)分别给出不同语言话者说相同内容话语/*bizning vygE qachan kilisen*/时基频实验结果。

表 2 维哈柯语言男声发话语音基频实验结果

语言对(男)	最大值	最小值	平均差	平均相似度
u-h	169.55/150.745	114.646/112.012	0.533	0.988
u-k	169.55/136.83	114.646/100.55	0.622	0.992
h-k	150.745/136.83	112.011/100.55	0.089	0.997

表 3 维哈柯语言女声发话语音基频实验结果

语言对(女)	最大值	最小值	平均差	平均相似度
u-h	259.04/291.45	184.44/150.46	0.469	0.995
u-k	259.04/326.82	184.44/200.01	0.027	0.994
h-k	291.45/326.82	150.46/200.01	0.442	0.997

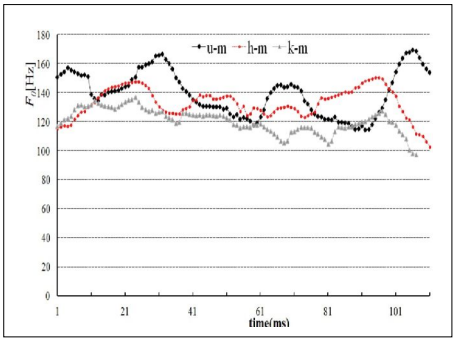

图-11(a) 维哈柯语言男声发话语音基频曲线比较

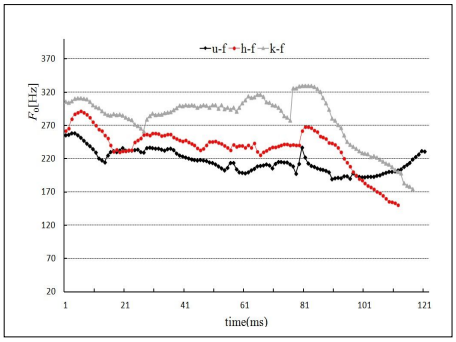

图-11(b) 维哈柯语言女声发话语音基频曲线比较

从表 2 和表 3 以及图-11(a)和图-11(b)观察到，维哈柯各语言发音风格几乎接近，在不同民族的男女发话中，h-k(哈柯)说话人音律最相似，其次是 u-k(维柯)，接下来是 u-h(维哈)。特别是，维语男声(u-m)有明显的音调特征。

5. 结论和讨论

大数据条件下建立共享云模型实现相似语言横向或者跨语言信息处理，多方位通信，促进少数民族语言的信息化发展是十分重要的研究方向。本文以阿勒泰语系下维哈柯及蒙古语为研究对象，利用平行语料从文本层面和发音层面研讨了同语系下诸多语言间的相似性或者相通性，定量给出这些语言间的相似程度。实验结果显示，在文本层面同族语言间以词为单元的文本转换的可能性较高；在语音层面维哈柯语言完全利用共享语音模型横向实现语音转换的可能性也较高。也就是说，如果在具备维吾尔语语言资源的前提下，通过横向处理方式实现哈语，柯语或者蒙古语多语种之间的机器转换，语音识别及语音合成等技术是完全有可能的，然而对于相似语言横向处理共享模型应该如何建设，还需要进一步研究。

参考文献

[1] Wushour Slam, *et al*, Speech Processing Technology of Uyghur Language[C]. Oriental COCOSDA International Conference on Speech Database and Assessments, 2009: 11-16.

[2] 卡哈尔江等, 一种改进的维吾尔语句子相似度计算方法[J], 中文信息学报,2011, Vol.25(4), 50-53.

[3] 伊·达瓦 等, 语料资源缺乏的连续语音识别方法的研究[J], 自动化学报, 2010, Vol.36(4), 550-557.

[4] Shuichi Itahashi and Chiu-yu Tseng, COMPUTER PROCESSING OF ORIENTAL LANGUAGES[M]. 2010, World Scientific,www.American-sGroup.com.

[5] T.Schultz and A.Waibel, Fast Bootstrapping of LVCSR System with Multilingual phoneme Sets[C], Proc. Eurospeech 1. 371-374.

[6] Lin jun Zhang, *et.al*, "Cross-Language information retrival", Journal of Computer Science,2004,Vol.31(7), 16-19.

[7] EHARA Terumasa, *et al*. "Mongolian to Japanese machine translation system C. Proceedings of second international symposium on information and language processing, 2007, pp.27-33.

[8] Idomucogiin Dawa, Satoshi Nakamura, "A Study on Cross Transformation of Mongolian Family Language", Journal of Natural Language Processing, J-STAGE, Vol.15 No.5,2008, pp3-21.

[9] 达瓦·伊德木草 ,"基于机器翻译的蒙文多文本转写方法的研究",新疆维吾尔自治区科技厅自然科学基金,项目编号为 2011211A012。项目起始时间 2011.6~2013.12.

[10] 伊·达瓦等, 蒙古语语言-文字的自动化处理[J]. 中文信息学报,2006, Vol.20(4), 56-62.

[11] Jun. Ye, "Cosine Similarity measures for intuitionistic fuzzy sets and their Applications". Mathmatical and Computer Modeling, 2011 Vol.53, 91-97.

[12] T.Schultz and A.Waibel,. EXPERIMENTS ON CROSS LANGUAGE ACOUSTIC MODELING[C]. 2001 EUROSPEECH.

[13] 古井 贞熙 著, 音响·音声工学[M], 东京, 近代科学社,1992.

[14] 伊·达瓦, 大川 茂村,白井 克彦, 蒙古语七个元音声频特性计算机分析[J], 声学学报 , 1999, Vol.24(1), 94-97.