THUYG-20：一个免费的维吾尔语语音数据库*

爱斯卡尔▪肉孜 殷实 张之勇 王东 爱斯卡尔▪阿朴杜拉 郑方

文 摘：语音数据资源是语音识别研究的基础。当前国内几乎没有开放的语音数据库供研究者免费使用，特别是在少数民族语音识别方面，数据资源更为贫乏。本文公开一个免费的维吾尔语连续语音数据库，该数据库包括约 20 小时的训练数据和 1 小时的测试数据。我们同时公开了构建维吾尔语语音识别系统所需要的音素集、词表、文本数据等相关资源，公开了用于构建基线系统的脚本，给出该基线系统在纯净测试数据和噪音测试数据上的识别性能。

关键词：维吾尔语；语料库；语音识别；DNN；

中图分类号： TP39

语音数据库是进行语音识别研究的基础。国际上比较著名的几个数据库包括 RM, Aurora, TIMIT, WSJ, Switch Board 等。这些标准数据库使得研究者可以在同一标准下比较自己的研究方法，因而极大推动了语音识别技术的发展。然而，这些数据库都需要付费才能使用，限制了初学者对语音领域的涉足。这一状况目前有所改变，出现了如 LibriSpeech 等可以免费使用的数据库资源[1]。

相对而言，国内的数据库标准化工作十分薄弱。到目前为止，国内用于语音识别研究的只有为数不多的几个数据库为研究者公认，如汉语 863 数据库。少数民族语言的数据库建设还处于空白阶段，仅有的几个数据资源由研究者独立录制，规模小，缺乏统一标准，普及面窄。特别是，这些有限的资源共享性差，免费数据资源几乎没有。这一现状极大制约了我国语音识别研究的发展。

本文公开一个约 20 小时的维吾尔语语音数据库（THUYG-20）供研究者免费下载使用。特别是，我们提供了用于构建完整语音识别系统的所有资源、代码、流程，提供完整的训练和测试标准，提供基线系统的识别结果，提供可重现的自动脚本。我们期望通过这一数据库和基线系统的公开，可以吸引更多对语音识别研究感兴趣的学者，促进学术交流，推动国内语音识别，特别是维吾尔语语音识别技术的发展。

本文结构如下：第 1 部分介绍维吾尔语语音数据库建设的相关研究成果，第 2 部分介绍 THUYG-20 数据库，第 3 部分介绍基于 THUYG-20 的基线系统构建和识别结果，最后在第 4 部分给出总结。

[1] http://www.openslr.org/12/

1 维吾尔语语音数据库研究现状

维吾尔语语音识别研究已经取得了一系列研究成果，如[1,2,3,5,6,11]。在研究过程中，学者们大多录制自己的数据库并在此基础上发表结果。本节对当前已有的数据库做一总结。

文献[1]中所使用的数据库由 171 个发音人组成，其中男性 85 人，女性 86 人，发音文本从前后 30 天的新疆日报（维文版）中选择的 1200 个句子。文献[2]中的数据库包括男女发音人各 10 人，1200 句用于训练，30 句用于测试。文献[3]用男女各 4 个人的语音数据进行训练，其余 2 个人用于测试。文献[4，16]中的数据库由音节、词语、语句、数字和常用符号等 5 个数据库组成。文献[5，19，21，22]用 356 人（189 女 167 男）128 小时的维吾尔语朗读式语音数据，选择男女各 5 个人的 1018 个语句（约 2 小时）用于测试。文献[6，7，8，18]中的数据库包括男女各 32 人的数据，每个发音人朗读 100 个随机选择的句子，其中 54 个人的数据用于训练，其余人的数据用于测试。文献[9，10]中的语料库为维吾尔语口语语料，包含词条分别为 21196 和 35056 条。文献[11]中的语料库由 1.2 万句约 9.6 个小时的语料组成。文献[12，13，23，24，25]中的训练集包含 353 个发音人的 150 个小时数据，测试集包含 23 个人的 1248 条语音数据。文献[14]中训练集约为 15 小时数据，测试集约为 0.5 小时数据。文献[15]中的训练集由 1052 个人的 470 个小时数据组成，测试集由 11 个人的 2186 条语句（约 2 个小时）组成。文献[20]中的数据库包含 94 个发音人，每个人参加 30 分钟左右的电话聊天。

可见，维吾尔语研究过程中确实积累了相当规模的语音数据。然而，这些数据由各研究机构内部或小范围合作者使用，数据库标准不统一，数据内容不公开，发表的研究结果无法由其他研究者重

现，也无法进行横向对比。更重要的是，由于各研究机构的封闭性，学者们在进行维吾尔语语音研究时多倾向于自行建立语料库，造成了严重的重复劳动和资源浪费，极大制约了维吾尔语语音识别技术的发展。因此，一个标准的、公开的、免费的、高质量的维吾尔语语音数据库，对推动维吾尔语语音识别及相关研究的发展具有重要意义。

2 THUYG-20:免费的维吾尔语语音数据库

本文发布维吾尔语语音数据库 THUYG-20 可以在网上免费下载[2]。数据库包含的资源包括：约 20 小时的语音数据，约 12M 单词的文本数据，包含约 4.5 万余单词的词表，基于 Kaldi 的系统构建脚本。本节我们给出该数据库的基本信息，下一节给出在两个测试集上我们得到的基线系统识别结果。

2.1 数据库规模

表 1 给出 THUYG-20 数据库中语音数据的统计结果，其中训练集用于声学模型训练，开发集用来选择模型参数，测试集用来进行性能测试。表 2 给出该数据库中文本数据的统计结果，其中训练集用来训练语言模型，测试集用来测试语言模型性能。

表 1 THUYG-20 语音语料库参数

语料库	说话人	男	女	年龄	句数	时长
训练集	348	163	185	19-28	7600	20.15
开发集	224	113	111	19-28	400	1.08
测试集	23	13	10	22-28	1468	2.4

表 2 THUYG -20 文本语料库参数

语料库	句子	单词	词素	音节	字符
训练集	1620k	11.58m	21.88m	31.74m	78.18m
测试集	11888	0.217m	0.408m	0.592m	1.46m

2.2 语音数据来源

1) 录音环境
   办公室环境，不包括其他说话人声音。
2) 录音设备
   IBM-联想台式机，外置麦克风。
3) 录音人
   348 名高校在校本科生及研究生，均为维吾尔族说话人，来自新疆 30 多个地州。
4) 录音内容
   常规话题，包括小说、报纸和各类书籍。
5) 录音时间：
   2012/1-2012/9

2.3 数据库用途

1) 维吾尔语语音识别研究
2) 维吾尔语说话人识别研究
3) 维吾尔语语音与语言特性研究

2.4 语料库数据规格

1) 信息文件：文本文件，存放说话人性别、年龄、族别及文化程度
2) 脚本文件：文本文件，存放说话人发音文本，每句文本包括句子编号、发音文本。
3) 语音文件：语音文件，存放说话人语音，文件名由性别、说话人编号和句子编号组成，如 F00108000148，表示编号为 108 的女性说话人，发音文本句子编号为 148。
4) 采样格式
   16KHz，16 位，单声道，wav 格式。

2.5 发布格式

网络发布，自由下载，对研究者免费。

2.6 测试任务

基于表 1 中的测试集，我们发布如下两个标准测试任务:

1) TEST-A: 测试数据由 THUYG-20 中的原始测试数据组成，即纯净无噪音数据。
2) TEST-N: 测试数据由 THUYG-20 中的原始测试数据混合一定比例的噪音组成。噪音来源为 DEMAND 噪音库[3]中包含的三种噪音：白噪音、汽车噪音和咖啡馆噪音，混合后的 SNR 包括从-6db 到 9db，共 18 个测试子任务。

3 基线维吾尔语识别系统构建

THUYG-20 提供的标准基线系统基于 Kaldi 开源工具包构建[30]。我们选择深度神经网络(DNN)作为声学模型，基于词的三元文法模型（3-gram）作为语言模型，基于 OpenFST[4]构建静态解码网络。

3.1 基于 DNN 的声学模型

DNN 是具有多隐藏层的神经网络。近年来一系列研究表明，DNN 比传统高斯混合模型（GMM）具有更强的声学建模能力。DNN 与隐马尔可夫模型(HMM) 结合的混合模型方法已经成为语音识别领域的主流框架。

图 1 为 THUYG-20 基线系统所采用的 DNN-HMM 模型框架。其中，DNN 模型的输入为基于 Mel 滤波器组的 Filter-bank（Fbank）特征，其中每帧语音长度为 25ms，帧移为 10ms，特征维数为

[2] http://cslt.riit.tsinghua.edu.cn:8081/data/thuyg20/README.html

[3] http://parole.loria.fr/DEMAND/

[4] http://www.openfst.org/twiki/bin/view/FST/WebHome

40。DNN 模型训练和解码时，首先以当前语音帧为中心，前后各取 5 帧组成上下文相关特征向量。这一特征向量经过线性判别式分析(LDA)映射为 200 维向量，再经过全局倒谱归一化(CMVN)去除信道影响后作为 DNN 的输入。

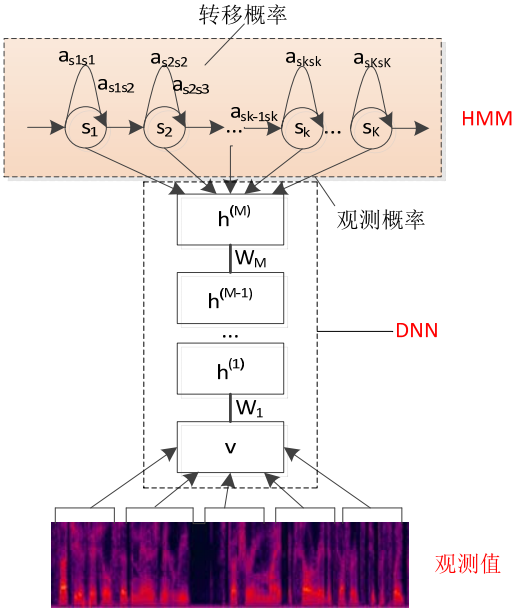

图 1 DNN-HMM 模型框架图

本文采用的 DNN 的结构如下：输入层含有 200 个输入单元，对应 200 维特征向量(Fbank+LDA+CMVN)；每个隐藏层含有 1200 个单元，共包含 4 个隐藏层； 输出层包括 3421 个单元，对应 HMM 系统中的 3421 个概率密度函数(PDF)。训练方法采用随机梯度下降(SGD)算法，训练准则包括交叉熵(xEnt)和最小音素错误率(MPE)两种。

3.2 语言模型

THUYG-20 基线系统采用基于词的 3-gram 模型作为语言模型。模型采用 SRILM 工具[5]进行训练，应用 KN-discount 平滑方法处理低频词和新词。这一模型训练方法简单，应用方便。

同时，我们报告基于词素(morpheme)的语言模型识别结果。该模型不作为 THUYG-20 基线系统的一部分，但提供了一种基于维吾尔语特性的语言模型增强方案。具本而言，因为维吾尔语具有很强的粘着性，词的形变数目众多，一方面对词表覆盖率提出了很大挑战，同时也产生了训练数据稀疏问题。文献[25] 中提出了基于词素对维吾尔语进行建模的方法，有效解决了上述问题。

具体而言，该方法首先对候选单词进行词干与后缀划分，然后选择最有代表性的词干-后缀二元组，以该二元组集合作为词表进行 3-gram 语言模型建模。通过将词降解为二元组，有效控制了词表规模，解决了数据稀疏问题，使得语言模型训练更加鲁棒。

3.3 加噪训练

THUYG-20 发布的测试数据集 TEST-N 包括混合各种比例噪声的带噪数据。我们采用[31]中提出的加噪训练方法提高基线系统在 TEST-N 上的识别性能。具体而言，在 DNN 训练过程中对训练数据随机加入多种噪声，使得 DNN 模型具有更好的可扩展性。实验证明这一方法可以极大提高基线系统的抗噪能力，且不会对纯静语音测试集(TEST-A)上的结果产生显著影响。

3.4 维吾尔语识别系统性能

本节给出我们基于 THUYG-20 构建的维吾尔语识别系统的识别性能。实验包括三组：第一组测试基线系统的识别结果，第二组测试基于词素语言模型的识别结果，第三组为引入带噪训练之后在 TEST-N 上的识别结果。

3.5 实验一：基线系统性能

THUYG-20 基线系统包括纯净语音训练的 DNN 声学模型，基于词 3-gram 的语言模型，基于 FST 的静态解码。测试包括基于交叉熵和基于 MPE 两种准则训练的 DNN 模型。识别性能采用词错误率(WER)进行评价,在纯净无噪声测试集 TEST-A 上的识别结果见表 3. 从结果上，利用 THUYG-20 数据库和基线系统构建流程，我们可以得到一个相对较好的维吾尔语识别系统。当然，20%左右的错误率和当前最好的英语、汉语等主流语言上的识别结果相比较还有相当差距，其中一个主要原因是 THUYG-20 的数据规模还不足以训练一个非常强大的声学模型。然而，我们相信通过增加训练数据来降低错误率并不是研究者关注的重点，基于当前数据规模（20 小时）采用更好的建模和训练方法以提高识别性能，比简单追求更大数据量对研究者更有意义。当前的性能指标只是提供一个基线标准，研究者可以在此基础上寻找更好的方法，特别是基于维吾尔语特性的优化方法，进一步提高系统性能。

需要说明的是，20 小时的数据规模与国际上广泛应用的 Aurora 4 数据库类似，这说明基于这一规模的数据库，研究者完全可以进行包括模型结构、模型训练方法、噪音去除、语音增强等多方面的研究。

表 3 基线系统在 TEST-A 上的识别结果

训练准则	WER（%）
交叉熵	19.57
MPE	18.95

[5] http://www.speech.sri.com/projects/srilm/

3.6 实验二：基于词素语言模型的系统性能

如前所述，基于 THUYG-20 可以进行声学和语言建模等多方面研究。文献[25]中提出的基于词素的语言模型建模即是利用维吾尔语词表特点在语言模型上进行的新探索。表 4 给出应用这一方法的识别结果。同实验一，测试在纯净数据集 TEST-A 上进行。可以看到，将词替换为词素进行语言模型建模可以有效提高系统的性能。

表 4　词素语言模型在 TEST-A 上的识别结果

训练准则	WER（%）
交叉熵	17.40
MPE	16.58

3.7 实验三：带噪训练系统性能

实验一和实验二中训练数据都是纯净数据，测试结果基于无噪声测试集 TEST-A。对于噪声数据集 TEST-N, 识别性能将显著下降。基于交叉熵训练的基线系统在 TEST-N 上的识别性能如表 5 所示，其中 white 代表加入白噪声的测试集，car 代表加入汽车噪声测试集，cafeteria 代表加入咖啡馆噪声的测试集,SNR 代表混入噪声后的测试数据信噪比。与表 3 中结果相比，加入噪声后，特别是白噪声和咖啡馆噪声，系统的识别系统显著下降。

表 5　基线系统在 TEST-N 上的识别结果

	WER (%)						
SNR(dB)	-6	-3	0	3	6	9	clean
white	99.96	99.85	99.37	96.19	86.82	72.76	19.57
car	23.42	22.17	21.15	20.56	20.27	19.95	19.57
cafeteria	97.96	91.80	79.67	63.57	49.25	38.02	19.57

为提高系统在 TEST-N 上的识别性能，我们采用加噪训练方法增强 DNN 模型(见 3.3 节)，加噪的具体参数见[31]。表 6 给出了系统经过加噪训练后的识别性能。可以看到，经过加噪训练后，在几乎所有测试条件下，系统识别性能都得到了显著提高。

表 6　基线系统加噪训练后在 TEST-N 上的识别结果

	WER (%)						
SNR(dB)	-6	-3	0	3	6	9	clean
white	76.35	62.75	51.61	41.56	34.91	30.45	19.67
car	21.85	21.04	20.52	20.14	20.00	19.84	19.67
cafeteria	66.47	51.82	40.02	31.96	26.97	24.15	19.67

4. 总结

本文发布了一个开放的维吾尔语语音数据库 THUYG-20，同时还发布了构建一个连续维吾尔语语音识别系统所需要的所有资源。我们希望通过这一数据库的公开，为在语音识别研究方面感兴趣的学者提供可以快速学习和切入的资源，为维吾尔语语音研究者提供可以进行对比验证的标准平台。我们公布了基于 THUYG-20 构建基线维吾尔语语音识别系统的方法，给出了该基线系统的性能，为维吾尔语语音识别研究提供了一个可以借鉴的标准。

值得说明的是，THUYG-20 不仅可用于语音识别研究，也可以用于说话人识别研究、维吾尔语语音和语言特性研究等多个领域。未来我们会进一步扩充该数据资源，并基于该资源发起更多合作研究和对比研究。

参 考 文 献

[1] 王昆仑， 樊志锦， 吐尔洪江等. 维吾尔语综合语音数据库系统[c]. 第五届全国人机语音通讯学术会议,中国哈尔滨，1998.

[2] 蔡琴,吾守尔·斯拉木. 基于 HTK 的维吾尔语连续数字语音识别[J]. 现代计算机，2007（4）

[3] 那斯尔江·吐尔逊，吾守尔·斯拉木，陶梅.基于 HTK 的维吾尔语连续语音识别研究[C]. 第七届中文信息处理国际会议,中国武汉，2007

[4] 王昆仑.维吾尔语音节语音识别与识别基元的研究[J].计算机科学,2003(7)

[5] 努尔麦麦提·尤鲁瓦斯,吾守尔·斯拉木,热依曼·吐尔逊. 基于音节的维吾尔语大词汇连续语音识别系统[J].清华大学学报,2013（6）

[6] Nasirjan Tursun, Wushour Silamu. Large Vocabulary Continuous Speech Recognition in Uyghur:Data Preparation and Experimental Results[C]. Chinese Spoken Language Processing, Kunming, China,2008

[7] Wushour Silamu, Nasirjan Tursun. HMM-Based Uyghur Continuous Speech Recognition System[C]. World Congress on Computer Science and Information Engineering, Los Angeles, USA 2009

[8] 那斯尔江·吐尔逊，吾守尔·斯拉木. 基于隐马尔可夫模型的维吾尔语连续语音识别系统[J].计算机应用，2009,29（2）

[9] 杨雅婷，马博，王磊等. 多发音字典在维吾尔语方言语音识别中的应用[J].清华大学学报，2011,51（9）

[10] 杨雅婷，马博，王磊等. 维吾尔语语音识别中发音变异现象[J].清华大学学报，2011,51（9）

[11] 张小燕，宿建军，薛化建等. 维吾尔语语音识别语料库中的ＯＯＶ研究[J].计算机工程与设计，2012,33（2）

[12] Mijit Ablimit, Graham Neubig, Masato Mimura. Uyghur Morpheme-based Language Models and ASR[C]. proceeding of ICSP,2010

[13] Mijit Ablimit, Askar Hamdulla , Tatsuya Kawahara. Morpheme Concatenation Approach in Language Modeling for Large-Vocabulary Uyghur Speech Recognition[C], Oriental COCOSDA, 2011

[14] 薛化建，董兴华，周喜等. 基于子字单元的维吾尔语语音识别研究[J].计算机工程，2011,37（20）

[15] Xin LI, Shang CAI, Jielin PAN. Large vocabulary Uyghur continuous speech recognition based on stems and suffixes, Chinese Spoken Language Processing (ISCSLP), 2010

[16] 王昆仑. 基于 CDCPM 的维吾尔语非特定人语音识别[J].计算机研究与发展，2001,38（10）
[17] Muhetaer Shadike, Xiao Li, Buheliqiguli Wasili. Large Vocabulary Continuous Speech Recognition of Uyghur: Basic Research of Decoder[C]. Advances in Neural Networks – ISNN,2011
[18] 陶梅,吾守尔·斯拉木,那斯尔江·吐尔逊. 基于 HTK 的维吾尔语连续语音声学建模[J]. 中文信息学报，2008,22（5）
[19] 努尔麦麦提·尤鲁瓦斯，吾守尔·斯拉木，热依曼·吐尔逊. 基于音节的维吾尔语大词汇连续语音识别系统[J].清华大学学报，2013,53（6）
[20] 米日古力·阿布都热素，艾克白尔·帕塔尔，艾斯卡尔·艾木都拉. 基于电话语料的维吾尔连续音素识别[J].通信技术，2012,45（7）
[21] 努尔麦麦提·尤鲁瓦斯，吾守尔·斯拉木，热依曼·吐尔逊. 维吾尔语大词汇语音识别系统识别单元研究[J].北京大学学报，2014,50（1）
[22] 努尔麦麦提·尤鲁瓦斯，吾守尔·斯拉木. 维吾尔语连续语音识别声学模型优化研究[J].计算机工程与应用，2013,49（2）
[23] Mijit Ablimit, Tatsuya Kawahara, Askar Hamdulla. Lexicon Optimization for Automatic Speech Recognition based on Discriminative Learning[C],APSIPA SC,xi'an,2011
[24] Mijit Ablimit, Tatsuya Kawahara, Askar Hamdulla. DISCRIMINATIVE APPROACH TO LEXICAL ENTRY SELECTION FOR AUTOMATIC SPEECH RECOGNITION OF AGGLUTINATIVE LANGUAGE[C],ICASSP,2012
[25] Mijit Ablimit, Tatsuya Kawahara, Askar Hamdulla. Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language[J]. Speech communication,2014(60)
[26] 武晓敏.基于 Julius 的维吾尔语连续语音识别研究[M].新疆大学，2012
[27] 王昆仑, 张贯虹, 吐尔洪江·阿布都克力木. 维吾尔语元音的声频特性分析和识别[J].中文信息学报，2010,24（2）
[28] 诺明花,吾守尔. 维吾尔语孤立词和连续数字语音识别设计与实现[C]. 第十一届全国民族语言文字信息学术研讨会,云南，2007
[29] 冯丽娟，吾守尔·斯拉木. 维吾尔语连续语音识别技术研究[J].研究与开发，现代计算机，2010（1）
[30] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C]，Proc of ASRU. 2011
[31] Shi Yin, Chao Liu, Zhiyong Zhang, Yiye Lin, Dong Wang, Javier Tejedor, Thomas Fang Zheng, Yinguo Li, Noisy Training for Deep Neural Networks in Speech Recognition[J], EURASIP Journal on Audio, Speech, and Music Processing 2015, 2015:2

THUYG-20：A Free Uyghur Speech Database

Abstract: Speech data plays a fundamental role in the research of speech recognition. At present, there is not yet an open speech database available for researchers in China, especially for minor languages such as Uyghur. This paper publishes a Uyghur speech database which is totally open and free. The database consists of 20 hours of training speech and 1 hour of test speech, as well as all the resources that are requested to construct a full-fledged Uyghur speech recognition system, including the phone set, lexicon, and text data. A recipe that is used to construct the baseline system is also published, and the results of the baseline system are reported on two test sets, one involves clean speech and the other involves noisy speech.

Key words: Uyghur language; corpus; speech recognition; DNN;