M2ASR 语音数据库设计方案

VERSION HISTORY

Version	Revision Author	Revision Date
0.1	李冠宇	2016.12.16
0.2	王东	2016.12.18

1. 语音数据库的要求

语音数据库对语音识别（ASR）系统的构造具有重要意义。在语音识别中，语音数据的主要作用是提供多场景的发音模式供机器学习，使得在后续的识别任务中该模式得以识别。因此，ASR 中一个合格的数据库应具各如下特点：

1) 广度覆盖性

一个大规模识别系统需要面对各种各样可能的识别场景，这些识别场景如果在训练语料中没有出现，则识别器在该场景中的识别效果将急剧下降。这意味着，为了构造一个好的识别系统，其训练数据中必须覆盖实际应用场景中的所有可能的环境和条件。这些环境和条件包括但不限于：（1）说话人（2）录音设备（3）传输信道（4）环境噪声（5）场景氛围（6）方式与情绪（7）地域口音（8）专业与领域等。对以上每种条件，都涵盖更为复杂的子条件，如地域口音中，可能包括轻微口音普通话，重口音普通话，地方方言，外国人汉语发音等。这些纷繁复杂的发音场景都会在不同程度上改变语音信号的实际形式，导致语音信号的高度变异性。这是语音识别困难的根源。

为对抗这种变动性，一般有两种方法，一是将识别任务规范到某一特定领域，从而减少语音信号的复杂性；二是收集更多的数据，让模型可以照顾上述各种变动性。我们首先考虑第二种方法，下节考虑第一种方法。

通常希望积累各种条件下的语音数据，使得模型可以覆盖全部上述条件与场景。这一朴素想法是合理的，在大多数条件下会较大幅提高系统在稀缺数据区的性能。然而，这一方法同时也会带来语音区分度的下降，因而降低系统在其它领域的性能。这一问题在传统 GMM 时代极为显著，因为大量数据会增加不同发音的 GMM 模型之间的交叠与混淆。神经网络时代，这一问题大为减弱，基于区分性建模和训练的神经网络方法可以在照顾覆盖性的同时避免混淆的产生。特别是进入深度神经网络(DNN)以后，语音特征可以由低层到高层分层次学习，因而将各种环境下的变动性逐层解释出来。换言之，DNN 将由环境不同而产生的各种变动分散表达成 DNN 中不同层次特征上的变动，在由低层到高层的逐层学习中逐步滤除，去掉变动，留下发音特性。因此，在 DNN 时代，增加语音数据产生的效果要远好于 GMM 时代，增加数据一般不会加重混淆，相反由于提供更多数据，模型训练变得更好。

从上述分析中，我们得知，所谓语音识别需要大数据，这个大数据并非单纯数据量大，而是覆盖面广。某一发音者 1000 小时的朗读数据远没有 100 发音者每人 1 小时的自由发音得到的 100 小时数据在实用中的效果好。积累数据最重要的是积累各种场景，积累各种发音条件和发音方式。

2) 领域深入性

积累大量广域数据有利于系统通用性的提高，但对某一应用领域没有太大效果。例如，我们有 100 小时朗读数据，如果增加 100 小时自由发音数据，则对自由发音则会产生大幅性能提高，但对朗读发音则提高不会明显。这意味着如果要提高某一

领域的性能，必须对这一领域进行特别的数据积累。这些领域相关数据也不应该直接与其它数据混合在一起，而是用自适应或转移学习的方法，在通用模型上进行自适应（adaptation）。

3) 选集区分性

语音数据的标注是非常费时费力的工作，因此需让标注工作产生最大价值。在一个数据集中，绝大部分数据对系统性能的贡献是有限的，特别是当基线系统比较强大时，绝大部分新数据并不能提供额外知识。真正有价值的数据是那些被当前系统识别错误的句子，这些句子提供了最有效的缺失信息。因此，当我们积累数据时，应尽量积累那些当前系统识别容易出错的数据进行标注。这意味着选择数据不仅考虑数据本身的价值，更需要考虑数据对系统带来的增量价值。

4) 标注准确性

ASR 对数据标注的要求在不同阶段是不一样的。一般来说，面对一个新的领域或任务的时候，我们希望快速积累领域数据，即使不精确也可以容忍。这时候标注准确性会让位于标注数据的数量，以得到一个更有效的识别模型为要。当系统较强大时，选集区分性的重要意义就体现出来，这时候就不该盲目收集和标注数据，而应该利用主动学习的方法，确定最有区分价值的数据进行标注。

基于上述数据需求原则，本文提出两类数据库的设计与标注原则：**人为设计数据库和实网语音数据库**。人为设计数据库(Human-designed database, HDD)指预先确定录音的文本内容，实网语音数据库(Real-life database, RLD)指由实际应用程序保存下来的实际场景语音数据。HDD 需组织人员录音，录音人员按预设文本进行录音。这一方式的录音成本较高，但后期效正成本较低; RLD 没有录音成本，但由于实网数据发音可能不清晰，标注成本较高。本文主要对这两类数据库的设计和标注提出基础方案。

2. 人为设计数据库(HDD)的设计方案

2.1 HDD 设计的基本原则

HDD 设计的主要问题在于：如何用最有限的录音成本实现最有效的语音现象覆盖。如前所述，语音现象包含的内容极为复杂：说话内容，说话人，录制设备，性别，口音，情绪，噪音，迟疑、连读、含混、吞音等口语现象。如此从多的语音现象指望一个人为设计的数据库全部覆盖是不可能的，特别是含混、吞音等口语现象很难由数据库设计实现。在有限的资源内，我们只能（1）覆盖有可能通过设计实现的语音现象; （2）覆盖尽可能多的语言现象。特别注意的是：

1） HDD 的设计不是为了一个实用系统，而是一个基线系统，是为了研究某一种语言的**基础语音现象，**因此通过复杂的 HDD 设计实现实用化，不仅是不现实的，也是不可取的。

2）实际系统最终会通过 RLD 才能实现，因此 HDD 不需考虑过多实际应用中的语音现象，如口语化、情绪 、噪音、连读等。再复杂的 HDD 设计，都不可能对这些口语化现象进行自然有效的覆盖。

3）将语音数据库设计和语言数据库设计区分开。语音数据库目的是为了覆盖语音现象，而非语言现象。因此，一些方言词汇、互联网词汇、口语表达等，不在语音数据库的设计范围之内，更加不在 HDD 设计范围之内。这些“语言”相关的现象需要收集相关文本，构造合适的语言模型实现。

基于上述基本原则，我们在设计 HDD 时应主要关注三个主要因素：（1）发音内容;(2)录音信道；（ 3 ）发音人。HDD 应尽可能覆盖这三个因素，对于其它因素可不予以考虑（在 RLD 中解决）。为此，我们希望：(1) 发音文本尽可能覆盖更多的音素和三音素；（2）信道尽可能多样化，如利用不同的手机；（3）发音人在性别、年龄分布上尽可能保持代表性和多样性。

2.2 HDD 设计基本步骤

1. 确定音素集

基于语音学家研究定义本语言的音素集。音素集的表示可多样化，如汉语的声韵母等。

2. 建立词典

（1）可能多的词汇，组成原始词表；

（2）依词条重要性排序。重要性可来源于在某一通用文本语料上的词频，也可来源于搜索引擎的搜索结果。

3. 词典发音标注

依排序后的词表得到若干大小的词典，进行发音标注。基础词表大约在 3 万左右，中等词表在 8 万左右，大规模连续识别系统在 12-20 万。这些发音标注需要人为检查。多发音情况需要标注所有可能发音。对一些发音规则较强的语言（如维语、哈语），可基于 grapheme 建模，因此不必进行发音标注。

4. 录音文本准备

如前所述，HDD 录音需要考虑（1）音素覆盖率 （2）信道 （3）说话人。为尽可能多的覆盖这些条件，录音时的这些条件应该是“交叉”的，即不同说话人应该考虑不同的信道，录制不同的句子。

因此，在选择录音文本时，应选择的文本句子数＝每人发音句子数 * 录音人数 * （1 - 重复系数）。其中重复系统代表同一句话出现在不同人录音文本中的可能性。

得到文本句子数后，从大文本数据库（种子数据库）中以 greedy search 的方法逐一挑选句子。在每一步挑选过程中，选择对 biphone 和 triphone 覆盖率增加最

大的句子，直到得到所需句子数的录音文本集。注意句子中的所有单词应可较方便得到发音（或者在词表中，或者基于 grapheme 建模）。生成的文本集中可加入特殊目的的句子（如领域相关句子或口语句子），但这些与语言模型更相关，如果获得比较困难，可在后续做语言模型中加入。

在生成每个人的录音文本时，如果重复系统等于 0，可均分录音文本集，否则可以用 sampling 方法，生成每个人的录音文本。

3. 实网语音数据库(RLD)的标注方案

RLD 与 HDD 完全不同，不是由数据库设计者设计，人为“故意”录制的，而是在用户不知情的条件下自由录制的，因此具有非富的口语现象和复杂的声学和语言学场景。采集 RLD 本身不是问题，关键是得到语音数据后的标注。本节描述 M2ASR 中所建议的 RLD 标注方案。

3.1. 标注总原则

“你觉得机器应该识别出什么，即标注什么”。

3.2 标注方法

- 成段语音用 VAD 工具切成句子。切分标准为非语音长度超过 0.5 秒，且每句话总长度不超过 5 秒。
- 所有汉语用简体汉字标注，包括数字，按实际发音标注。如“一二三”，而不是“123”。注意区分“一”和“幺”。句子中英文缩写部分按大写英文单词标注，如”IBM”，非缩写用小写英文单词标注，如”iphone”。
- 能理解的发音需标注成正确的汉字，不能理解的部分(如 “呃 啊 嗯 哦 唉”等)可用同音字代替。
- 问句标注问号”?”
- 转写内容要尽量与实际发音一致，不得删减。如发音为：“我是嗯北北京人”，应忠于原发音，而不是去掉语气词和重复，变成“我是北京人“。
- 明显非人声（铃声、叮声、咳嗽、扑话筒、音乐）、不可分辨的语音皆标为噪音”#”。长时噪音标一个”#”即可。长噪音中包含的若干可分辨的孤立词不必标注。
- 当发生混音的情况，如果可分辨主说话人(即某一说话人为清晰可辨的主讲，其音量尽大声音持续较长)，则尽量按主说话人进行标识。如果无法分辨出主体说话人，则混音部分标成噪音”#”。
- 允许标注人员将整句话标成全部噪音，这些句子包括：
  - ✓ 非本族语发音

- ✓ 整句纯噪音或乐音，或语音被噪音或乐音淹没
- ✓ 整句语音混杂，难以辨识

3.3 质量检查

1. 对训练数据，要求 5%复查结果的文字标注错误率在 5%以下。对测试数据，要求 100%复查，复查时对文字标注错误率达 1%以下方为合格。

2. 文字标注错误表示不可容忍的明显错误。对于因声音混杂，噪音过大，语音过快等造成的不可确定标注，不应计为错误。

3.4 标注样例

20152322 我想知道#是否合理#

20153377 #您能您能嗯#帮我查一下吗?

20157624 我想查一查一下 IBM 的电话

20157782 一个 iphone5 卖五百元吗?

4. 实际案例分析：藏语 HDD 数据库设计

依第 2 节所规定的 HDD 设计原则，我们设计藏语 HDD 数据库。我们的目的是采集 50 小时 100 人的手机 16k 录音。数据库从设计到录音应遵循如下步骤：

1) 文本正规化

藏语是一种拼音文字，但书写时具有较复杂的平面性。不同的编码方式对这一平面性进行不同处理，因此，一个重要工作是对藏语各种编码方式进行统一，转换成 Unicode 编码；

2) 分词

藏语以“字”为单位书写，而字与字之间没有明确分隔符。为增加模型建模能力，需要对字序列进行分词，转换成词序列。分词需要一个较大的初始词表，这一词表可由汉藏词典得到。

3) 词表选择

选择高频常用词 5 万作为基础词表，并依选定的音素集对该词表进行标音。

4) 种子语料收集

积累大规模文本语料（越多越好，5 万句以上）。该文本语料应保持多元化，覆盖较广的领域，最好包含一些口语表达。利用基础词表对种子文本进行分词。

5) 发音文本选择

每句长度在 5-25 个音节之间，平均时长以 4 秒计算，50 小时录音需要 4.5 万句，以 100 人录音计，每人需录音 450 句。我们选择 3 万句发音的话，每句话的复选率为 33.3%，这是可以接受的。

发音文本选择需注意如下规则：

a) 应尽量多地覆盖 triphone。

b) 句子长度在 5-25 个音节之间；

c) 剔除包含 OOV 的句子；

6) 语音采集

发音文本集确认后，对每个发音人随机生成至少 450 句发音文本，发送到发音者的手机端 APP，在线进行语音采集。

5. 参考文献

- 计算机藏文编码概述， 于洪志，1999，《西北民族大学学报(自然科学版)》, 1999(3):15-19
- 《藏汉大辞典》，张怡荪，1985 年民族出版社。