语音识别芯片 LD3320 高阶秘籍

Update@2011 年 04 月 20 日

语音识别芯片/声控芯片
单芯片/非特定人/动态编辑识别列表
语音识别解决方案

用声音去沟通
VUI (Voice User Interface)

ICRoute 用声音去沟通
VUI (Voice User Interface)

Web : www.icroute.com
Tel : 021-68546025
Mail: info@icroute.com

简介：.....3
一、 在识别精度要求高的场景中，使用“触发识别”模式.....3
二、 增添“垃圾关键词语”——吸收错误识别.....3
三、 口令触发模式.....4
四、 巧妙运用关键词语的 ID.....5
五、 工作电压.....5
六、 用拼音标注外文或者方言.....5
七、 对于同一关键词 ID 设置多个习惯发音.....6
八、 调节语音结束后得到识别结果的反应时间.....6
九、 麦克风，相关寄存器设置与识别效果和距离.....7
十、 语音识别用户使用模式详析.....9

简介：

基于语音识别芯片 LD3320 开发产品，可参考《LD3320 开发手册》。为了提高终端用户对于语音识别的主观体验，本文总结了一些高阶的方法和窍门，集结成文，与大家共享。

本文档会不定期更新，及时汇总实战中的经验。敬请期待。

一、 在识别精度要求高的场景中，使用“触发识别”模式

关于 LD3320 的两种使用模式，可以参考网站介绍：http://www.icroute.com/web_cn/LD332X_UserModel.html 。

在识别精度要求高的场景中，应该采用“触发识别”模式。原因是：

1） 用户在每次按热键后，精神处于最集中的状态，此时用户说的语音命令会比较认真，清晰。避免了用户过于随意的发音导致的识别误差。

2） 每次按热键后，产品应该给以一个明显的开始信号，比如发出“当”的一声或者其他提示信号，可以给用户一个明确开始的提示，方便用户掌握说语音命令的时间。

3） 由于按键触发后，用户就会贴近麦克风并说出语音命令，避免了其他环境声音被录入 LD3320 芯片导致的误识别。

另：这种方式还是一种省电的方式，在不识别时，彻底不让芯片工作以省电。

二、 增添“垃圾关键词语”——吸收错误识别

在设定好要识别的关键词语后，为了进一步降低误识别率，可以再添加一些其他的任意词汇进识别列表，用来吸收错误识别，从而达到降低误识别率的目的。

可以把这些关键词语称之为“垃圾关键词语”。

比如，某个应用场景中，需要识别的关键词语是 4 条，“前进”，“后退”，“开门”，“关门”。在把这 4 个关键词语设置进 LD3320 后，可以再另外设置 10~30 个词语进 LD3320，比如“前门”，“后门”，“阿阿阿”，“呜呜”等等。

只有识别结果是 4 个关键词语之内的，才认为识别有效。如果识别结果是“垃圾关键词语”，则说明是其他的声音导致的误识别，产品应该重新开始一次识别过程。

这样，可以非常非常有效地降低误识别率。极大地提高终端用户的主观使用体验。

“垃圾关键词语”的选取，最好可以选择一些字数和关键词语一样的词语，用来吸收可能发生的错误识别。

需要说明的是：这一方法，即可以应用在“触发识别”模式中，也可以应用在“循环识别”模式中。

这样作的原理如下：

非特定人语音识别技术 ASR，是一个基于关键词语列表的匹配识别技术，算法本质是在提取输入声音的特征后，在关键词语列表中寻找一个相似度最高的词语作为识别结果。(http://www.icroute.com/web_cn/LD332X_principle.html)

因此，任何的声音输入进语音识别芯片，都会去和关键词语列表中的词语进行匹配对比，并且也都会依次打分。这样，其他人在随意聊天，或者任意说一个不在关键词语列表中的命令，或者是其他毫无联系的说话声音，都可能会匹配到某一个关键词语并作为结果输出。从而造成误识别。

虽然算法设计中有一定的算法来避免出现这样的误识别，但还是不可完全避免。产品开发者可以在芯片外部针对性的处理以降低误识别率。本节提供的方法，是非常有效的一种方法，在实际应用中具有非常重要的地位。

三、 口令触发模式

在一些应用场合，希望识别精度高，但是又无法要求用户每次都用手按键来“触发识别”。此时，可以采用“口令触发模式”。

产品定义一句短语，作为触发口令。比如，可以定义“芝麻开门”作为触发口令。

产品在等待用户触发时，启动一个“循环识别”模式，把触发口令“芝麻开门”和其他几十个用来吸收错误的词汇设置进 LD3320。只有当检测到识别出的结果是触发口令时，才认为是终端用户叫了这个口令。此时，给出提示音，并启动一个“触发识别模式”，并且把相应的识别列表设置进 LD3320，提示用户在提示音后几秒钟内说出要执行的操作。

在等待用户的过程时，如果识别的结果是那些用来吸收错误的词汇，则认为是误识别，或者其他的声音干扰，而不进行任何的处理，直接再次进入“循环识别”模式。

这种口令触发模式，融合了其他两种模式的优点，并且结合第二节提到的“垃圾关键词语”的方法，可以为产品提供更加方便实用的语音操作特性。

四、 巧妙运用关键词语的 ID

在把关键词语设置进 LD3320 时，是把关键词语的拼音串传入 LD3320，并同时传入一个 ID，用来代表这个关键词语。

LD3320 的识别结果，也是把识别出的关键词语的 ID 作为结果输出。

在 LD3320 芯片，不同的关键词语是可以对应同一个 ID。而且 ID 不需要是连续的。这样就为产品开发者提供了很方便的编程手段。

例如：“北京”，“首都”，可以设置为同一个 ID，进行后续处理。

例如：在使用第二节提到的“垃圾关键词语”时，可以把添加的这些用来吸收错误的关键词语的 ID 都标记成一个值，或者把它们标注为比较特殊的 ID 值，如大于 200。在程序中就比较简单，很容易处理误识别，避免了添加进很多关键词语后，写程序中需要为这些关键词语的处理增加过多的程序分支。

五、 工作电压

LD3320 有三路电源输入，

VDD 数字电路用电源输入 3.0 V - 3.3 V

VDDIO 数字 I/O 电路用电源输入 1.65 V - VDD

VDDA 模拟电路用电源输入 3.0 V - 4.0 V

但是在实际设计时，可以采用统一送入 3.3v 的工作电压给这三路电源。最低工作电压是 3.0v，当输入电压低于这个数值时，芯片会无法启动工作。
这样可以简化电路设计。

如果条件允许，可以把模拟电源和数字电源隔离开，避免干扰，以取得最佳的电源管理效果。

六、 用拼音标注外文或者方言

语音识别，识别的是“语音”。对于非特定人语音识别来说，在描述关键词语时，是用音标标注出要识别的关键词语。

对于目前 LD3320 支持的中文识别来说，就是用拼音来描述出关键词语。

也就是说，只要是拼音可以拼出的发音，都是可以输入芯片并进行识别的。

因此，在某些场合需要识别一些简单的外文或者纯方言发音的时候，可以用拼音标注的方法来实现。

例如，有些场合需要识别一些简单的英文单词，可以用拼音标注：

one -→ wan

two -→ tu

three -→ si rui

例如，有些场合需要识别一些纯方言发音的词汇，也可以用拼音标注：

上海话的“晚”发音是“ya”，那么“晚报”这个词汇，用普通话标注是“wan bao”，如果要标注成上海话发音，就是“ya bao”，这样上海话说的“晚报”也就可以被识别了。

值得注意的是：LD3320 支持的是中文普通话，有些外文或者方言发音无法用拼音描述，所以 LD3320 不一定能够完成所有需要的外文或者方言任务。

七、 对于同一关键词 ID 设置多个习惯发音

终端用户在说语音指令时，可能对同一个词汇有不同的发音习惯。

例如，“打开电灯”，用户可能会说“开灯”，“打开灯”，“打开电灯”，“把灯打开”等等。

充分利用 LD3320 的 50 条可动态编辑的关键识别条目的特性，开发者可以把这些习惯发音都设置进芯片，这样无论用户怎么说，都会被正确识别出来，进一步增加终端用户的良好体验。

同时，可以结合第四条秘籍“巧妙利用关键词语的 ID”，在编程中可以很方便地处理这些多个习惯发音。

值得注意的是：如果用来进行控制工作，需要加入一些垃圾关键词吸收错误以降低误识率。见第二节“增添“垃圾关键词语”——吸收错误识别”。

八、 调节语音结束后得到识别结果的反应时间

LD3320 芯片内部是通过 VAD（端点检测）机制来判断人是否说完语音，并给出识别结果。关于 VAD 以及获得识别结果机制的详细说明，请阅读网页的介绍：http://www.icroute.com/web_cn/LD332X_principle.html

根据 VAD 机制，语音识别芯片监测出有一段连续的背景噪音后，认为用户已经说完了语音识别命令，然后再给出识别结果。

默认设置是监测到在人声开始后有连续的 600 毫秒的不说话时，才会给出识别结果。

也就是说，根据默认设置，从人说话结束，到语音识别芯片主动送出结果中断，至少要有 600 毫秒的间隔，如果用户希望调节这个反应间隔，可以从以下几方面入手：

1. 改变使用方式

采用类似于步话机的方式，每次人按键后，按下不放，开始说命令，说完命令后，松开按键，每次检测到松开按键时，主控的单片机通过设置 BC 寄存器来立即获得识别结果。（BC 寄存器见“开发手册”的说明）

2. 修改 VAD 判断的寄存器

B5 寄存器

ASR: Vad Silence End 在语音检测到语音数据段以后，又检测到了背景噪音段，连续检测到多长时间的背景噪音段才可以确认为是真正的语音结束。每 1 单位，10 毫秒。Default：60，相当于 600 毫秒数值范围：20~200 (相当于 200~2000 毫秒)

但是这个修改会导致，如果这个时间过短，导致用户在说话时的说话停顿也会造成 VAD 检测认为说话结束，从而降低某些用户的识别率。

3. 修改麦克风音量寄存器

修改麦克风的音量，35 寄存器，（建议调整范围在 40H~58H 之间），看哪个录音增益适合使用的麦克风，以及使用的环境。

4. 修改 B8 寄存器

修改 B8 寄存器。

比如修改为 2， 那么这意味着，无论如何，在每次识别开始后 2 秒钟的时间内，必然会停止识别给出一个识别结果。（这个设置不影响 VAD 检测）。

如果 b8 值特别小，比如设置：1，2，3，就需要在开始识别前，给用户一个很明确的提示，要开始识别了。免得用户还没有准备就识别时间过去了。

但这个间隔设置的过短，也必然会引起一些可能存在的误识别，比如语音命令比较长，那么这个时间设置的太小，就会造成比较长的语音命令无法在特定时间内完整念完引起误识别。

所以当这个数值设置比较小的时候，一般建议使用“触发识别”的用户界面，避免使用“循环识别”的用户界面。

5. 使用环境

改变使用环境，或许在某些环境中的噪声或者回声会影响到判断说话结束。

以及说话人自己的音量，如果声音很低，也会导致判断人说话是否结束比较困难。

改变命令词语内容，比较好念，开口音响亮等，方便使用者连续清晰念出语音命令。

九、 麦克风，相关寄存器设置与识别效果和距离

语音识别的效果，是一个主观体验的结果。和以下的因素都有关系：

1. 周围环境的声音
2. 识别列表的内容设置：是发音响亮的开口音还是不容易发音的闭口音
3. 识别列表各个词语之间的相互差别程度
4. 说话人的发音清晰/大小/快慢/认真程度/口音
5. 用户操作流程的设置
6. 外接麦克风的物理特性

7.说话人是否放开音量
等等。

语音识别的质量和有效作用距离和麦克风/咪头的关系非常大。麦克风/咪头的质量决定了送怎样的声音质量给识别芯片，所以也决定了识别的效果的距离。一般地咪头作用距离大概在 1 米左右，仔细挑选的咪头作用距离大概在 2~3 米。

取决于生产的工艺和质量，导致麦克风/咪头的背景电噪声高低， 从而导致不同麦克风/咪头的录音距离不一样。比如同样号称 39db 灵敏度的咪头，从市场上不同的地方买回来，效果完全不一样。

质量差的，录音的电噪声非常高，导致会把人的说话声音湮没，需要人提高音量，或者距离凑近。

质量好的，录音的频响曲线比较平整，电噪声低，就可以把比较远的人声比较清晰地录入。

市场上有很多咪头，是针对手机生产的，他们的特点就是近距离录音。这样在手机上应用是可以压抑远处的噪声。但是在语音识别应用上，这样的近距离咪头会造成识别效果下降，识别距离近等结果。

开发者应该多实验一些麦克风/咪头来选取适合自己产品的。需要根据产品的应用环境以及定位，选用合适的麦克风，来充分发挥 LD3320 识别芯片的识别功效。

值得注意的是：对于加了放大器来扩大识别距离的麦克风或者类似装置，由于某些放大器会破坏声音的波形，造成声音输入的“过冲击”造成声音严重失真，会极大地影响识别效果，导致识别率严重下降。

值得注意 2：对于加了降噪芯片的模块，由于某些数字降噪芯片会把一些声音比较细小的声音强行归置为 0，导致丢失声音（比如发音比较轻的辅音“si”，“wu”等），也会极大地影响识别效果，导致识别率严重下降。

在调节与麦克风 AD 输入相关的寄存器时，有以下一些建议，可以有助于提高识别距离和提升识别质量：

1. **MIC 增益寄存器 0x35**

LD3320 芯片的 mic 增益寄存器(0x35)，并不是设置得越大约好，对于 35 寄存器，一般建议范围在 0x40 ~0x53 之间。 再高的数值，会导致非常容易出现过激，识别率严重下降。参考程序给的是 0x43。 如果需要调高录音音量，建议到 0x4c，就应该比较合适，

2. **B3 寄存器**

寄存器 0xB3，大概可以理解为是灵敏度，就是对周围声音的强度的反应灵敏度。默认是 0x12H。

在需要增加灵敏度，扩大识别距离时，调节到 0x0FH 或者 0x0AH 实现效果。 （最灵敏是 0x1）

但是这个调整是双刃剑。对声音敏感了，必然会带来识别率上的负影响。所以对于声音的调整是非常需要平衡。需要结合产品慢慢调整。

可以在设置 0x35 麦克风音量的地方去设置这个寄存器到 0x0FH。

3 在按键触发模式下可以关闭 VAD

如果产品可以使用按键触发方式，就是在按键后，定时录音一段时间，（比如 3 秒钟，或者 5 秒钟）然后在这段时间结束后再得识别结果。

那么有两种选择，一种是还是保持 VAD 打开，这样还是会先检测人说话是否开始，在检测到人说话结束后，就会给出中断和识别结果。另一种是彻底关闭 VAD。把这段时间（比如 3 秒钟）的所有声音数据都进行识别运算。只有在这段时间结束后，才会给出中断和识别结果。

这种按键触发并且关闭 VAD 的方式，寄存器按如下设置：0xB3=0 0xB8=2(时间长度，单位秒)。当把 0xB3 寄存器设置为 0 时，相当于关闭了 VAD 检测。此时芯片会把送入的所有声音都进行识别运算，而不再检查和区分人声和背景噪声。设置的 B8 的数值，决定了固定录音的时间长度，比如设定为 2 或者 3 ，等数值，结合产品的实际需求来设定。

4. B5 寄存器

B5 寄存器，（见第八节），默认是 60，相当于 600 毫秒。如果这个被设置的过短，会导致人说话中间的某些发音比较轻的辅音，会被当作说话已经结束，从而导致长字词被分开，从而导致识别错误。

5.

综上，如果想把声音录入的比较大，识别作用距离远一些，在寄存器的调整上，可以把 0x35 寄存器设置为 0x4c 左右，0xB3 寄存器设置为 0xf 左右， B5 B8 采用默认。 同时，再选用一款质量优秀的适合的麦克风，二者结合起来，会达到比较好的识别效果。

十、 语音识别用户使用模式详析

在网站页面 http://www.icroute.com/web_cn/LD332X_UserModel.html 介绍了两种不同的用户使用模式：触发识别模式和循环识别模式。

在下载页面的技术文档：《**声控智能产品语音界面设计指南.pdf**.》（http://www.icroute.com/web_cn/VUI_DesignManual.html ）和本文档的前面几节，也多次提到了在不同场合，需要使用不同的用户模式。

其实，如果从芯片的技术角度来看，这些不同的用户模式，语音识别芯片 LD3320 的工作流程都是一样的：
芯片**初始化** LD_Init_ASR()—〉**添加关键词语**识别列表 LD_AsrAddFixed()—〉打开麦克风并**启动语音识别** LD_AsrRun()。（这个流程，也就是函数 RunASR()的内容）

启动这个识别流程和结束识别流程的条件不一样，就构成了不同的用户模式。

结束识别流程的条件有：
1） 在打开 VAD 的情况下，VAD 检测到有声音结束（有一段持续的静音出现），则识别流程结束，给出中断。
2） B8 寄存器设置的时间到了，如果仍在识别流程中进行识别运算，则识别流程结束，给出中断。
3） 向 BC 寄存器写入 07H 或者 08H，强制结束识别流程，给出中断。

这三个条件是互相不依赖的，任何一个条件先达到，都会结束识别流程。有关 VAD 寄存器的详细说明请阅读《**LD332X 开发手册.pdf**》http://www.icroute.com/web_cn/Download.html#LD332X-Manual ； 以及本文档第八，第九节的相关讨论。

开始识别流程的条件是主控 MCU 控制的。

把各种开始和结束识别流程的条件组合起来，就形成了多种不同的用户使用模式，开发者可以根据自己产品的特点，来选择最合适的。这些用户使用模式如下：

循环识别：

- 打开 VAD 功能（默认）。
- 设置 B8 寄存器为合适数值（默认）：一般建议这个时间比较长，避免在循环识别过程中，用户的说话正好落被两次识别流程分割开。
- 识别流程开始条件：一旦有识别中断，则立即读取识别结果，处理完毕后，立即启动下一次识别流程。
- 识别流程结束条件：由 VAD 结束或者由 B8 条件结束。

触发识别+单按键+VAD：

- 打开 VAD 功能（默认）。
- 设置 B8 寄存器为合适数值：时间由用户根据产品特点设定，此处用意为按键后最多等待多长时间，过了这个时间，就不再接收用户的语音命令。
- 识别流程开始条件：主控 MCU 接收到按键，则启动一次识别流程
- 识别流程结束条件：由 VAD 结束或者由 B8 条件结束。

触发识别+单按键+VAD+避免零结果：

这个模式是对上一种模式（触发识别+单按键+VAD）的改进，由于不同的麦克风的灵敏度完全不一样，环境噪声也不一样，所以即使人不说话，也有可能由于其他原因（其他一个声音进入麦克风引起 VAD 检测到声音开始）导致 VAD 启动并结束一次识别流程。此时识别结果一般为零。而此时往往用户还没有开始说话。为了避免这种情况，可以作改进。

- 打开 VAD 功能（默认）。
- 设置 B8 寄存器为合适数值：时间由用户根据产品特点设定，此处用意为按键后最多等待多长时间，过了这个时间，就不再接收用户的语音命令。
- 识别流程开始条件：主控 MCU 接收到按键，则启动一次识别流程。当识别结果为零结果时，则立即再启动一次识别流程。直到某次识别流程给出一个识别结果。
- 识别流程结束条件：由 VAD 结束或者由 B8 条件结束。

补充说明：当关键词语列表设置了垃圾词语的时候，此时开发者需要自行决定，是只要给出一个非零的识别结果就结束（如果识别出来是垃圾词语，则本次按键触发的最终识别结果就是垃圾词语）；还是一定要识别到一个非垃圾词语的结果才结束（有可能会出现用户一直不说话或者一直不认真说，导致识别不出来有效识别结果，而一直在识别）

触发识别+单按键+VAD+避免零结果+主控 MCU 定时：

针对上一种模式，还可以再作改进，就是引入主控 MCU 的定时功能。主控 MCU 处也起一个 Timer 来定时，一旦这个时间到，如果还处于识别流程过程中，则立即向 BC 寄存器设置 07H 或者 08H 来强制结束识别。同时，主控 MCU 也不再重新启动识别流程。

- 打开 VAD 功能（默认）。
- 设置 B8 寄存器为合适数值：时间由用户根据产品特点设定，此处用意为按键后最多等待多长时间，过了这个时间，就不再接收用户的语音命令。
- 识别流程开始条件：主控 MCU 接收到按键，则启动一次识别流程。当识别结果为零结果时，则立即再启动一次识别流程。直到某次识别流程给出一个识别结果。
- 识别流程结束条件：由 VAD 结束或者由 B8 条件结束或者主控 MCU 定时到后主动结束。

补充说明：在主控 MCU 定时到后，想 BC 寄存器设置 07H 还是 08H，需要开发者自行决定，是否认为这样定时结束时拿到的识别结果会有效果。

触发识别+单按键+关闭 VAD：

对于单按键的触发模式，还可以关闭 VAD，只依靠 B8 的时间设置来结束识别流程。这种方式，和手机上语音王的方式是一样的，就是在按键后，一定需要过完设定的时间后，才给出识别结果。这种情况下，是把这段时间内的所有声音都送入识别芯片进行运算，如果使用者配合，识别率在理论上应该比启动 VAD 要好一点，但如果使用者不配合，可能会差一些。

- 关闭 VAD 功能。
- 设置 B8 寄存器为合适数值：时间由用户根据产品特点设定。一般建议设置为 3~5 秒比较合适
- 识别流程开始条件：主控 MCU 接收到按键，则启动一次识别流程。
- 识别流程结束条件：由 B8 条件结束。而且每次必定是过完 B8 设置的时间后，才会给出识别结果，不会提前给出。即使用户说完命令，如果不到 B8 设定的时间，也不会给出识别结果。

触发识别+双按键+关闭 VAD：

双按键的模式类似于步话机模式，就是按一下按键后，开始说语音命令，当说话语音命令后，再次按下按键，获得识别结果。需要使用者自己在说话命令后按键通知 MCU。

- 关闭 VAD 功能。
- 设置 B8 寄存器为合适数值：B8 一般建议设置比较长一些，因为这种模式是由使用者来控制结束识别流程，所以应该避免 B8 时间过多而提前结束识别流程。
- 识别流程开始条件：主控 MCU 接收到按键，则启动一次识别流程。
- 识别流程结束条件：主控 MCU 再次接收到按键，向 BC 寄存器设置 07H，结束识别 流程并获得识别结果。

补充说明：如果使用者肯配合，那么这种模式应是效果最好的。因为使用者通过两次按键，严格地只把语音命令送入芯片，而不把背景噪音送入芯片。所以效果为最好。

当然，开发者也可以修改为，按下按键后开始说话，松开按键后给出识别结果。这个是在主控 MCU 中的开发工作了。

以上，是详细说明了几种不同的语音使用模式，都是由开发者在主控 MCU 中编写代码来实现，需要开发者结合自己的产品实际特点，来精心选择最合适的方式。同时还需要告诉使用者如果正确使用，在产品设计中给使用者以明确的提示音或者提示灯光来帮助使用者在正确的时间内说出语音命令，获得最佳的识别体验。