[知乎](https://www.zhihu.com/)

* [注册知乎](javascript:;)
* [登录](javascript:;)

提问

窗体顶端

知乎搜索搜索

窗体底端

* [首页](https://www.zhihu.com/)
* [话题](https://www.zhihu.com/topic)
* [发现](https://www.zhihu.com/explore)

[人工智能](https://www.zhihu.com/topic/19551275)[科普](https://www.zhihu.com/topic/19551585)[语音识别](https://www.zhihu.com/topic/19560846)[X 的原理](https://www.zhihu.com/topic/19561178)[X 的工作原理](https://www.zhihu.com/topic/19566024)

**语音识别的技术原理是什么？**

[1 条评论](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

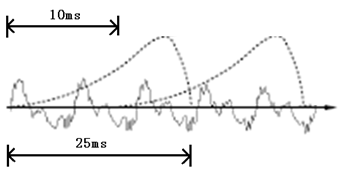
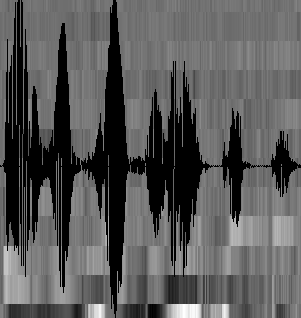
按投票排序[按时间排序](https://www.zhihu.com/question/20398418?sort=created)

**11 个回答**

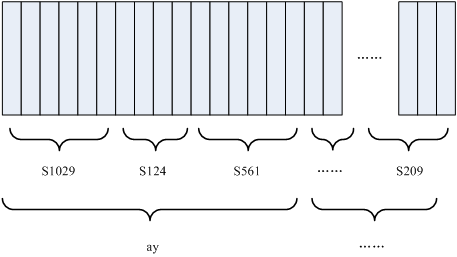
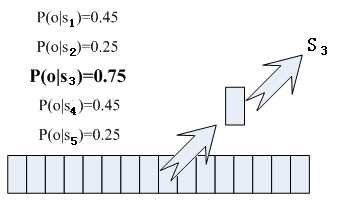
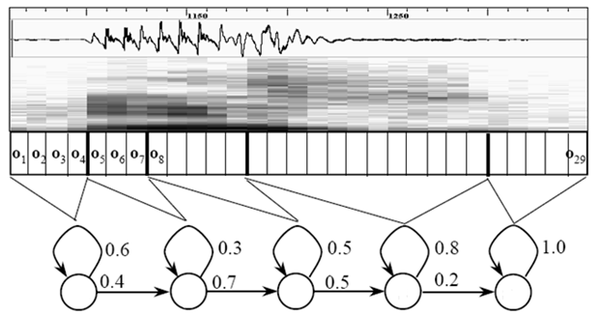
1495赞同反对，不会显示你的姓名

[https://pic1.zhimg.com/1be9d5c0ccc6b73c724ed50812dc2350_s.jpg](https://www.zhihu.com/people/jimbozhang)[**张俊博**](https://www.zhihu.com/people/jimbozhang)乐乐语音

收录于 [知乎周刊](https://www.zhihu.com/publications/weekly/19550919) 、 编辑推荐 •[1495 人赞同](https://www.zhihu.com/question/20398418)

简要给大家介绍一下语音怎么变文字的吧。**需要说明的是，这篇文章为了易读性而牺牲了严谨性，因此文中的很多表述实际上是不准确的。**对于有兴趣深入了解的同学，本文的末尾推荐了几份进阶阅读材料。下面我们开始。  
  
首先，我们知道声音实际上是一种波。常见的mp3、wmv等格式都是压缩格式，必须转成非压缩的纯波形文件来处理，比如Windows PCM文件，也就是俗称的wav文件。wav文件里存储的除了一个文件头以外，就是声音波形的一个个点了。下图是一个波形的示例。  
  
  
在开始语音识别之前，有时需要把首尾端的静音切除，降低对后续步骤造成的干扰。这个静音切除的操作一般称为VAD，需要用到信号处理的一些技术。  
  
要对声音进行分析，需要对声音分帧，也就是把声音切开成一小段一小段，每小段称为一帧。分帧操作一般不是简单的切开，而是使用移动窗函数来实现，这里不详述。帧与帧之间一般是有交叠的，就像下图这样：  
图中，每帧的长度为25毫秒，每两帧之间有25-10=15毫秒的交叠。我们称为以帧长25ms、帧移10ms分帧。  
  
分帧后，语音就变成了很多小段。但波形在时域上几乎没有描述能力，因此必须将波形作变换。常见的一种变换方法是提取MFCC特征，根据人耳的生理特性，把每一帧波形变成一个多维向量，可以简单地理解为这个向量包含了这帧语音的内容信息。这个过程叫做声学特征提取。实际应用中，这一步有很多细节，声学特征也不止有MFCC这一种，具体这里不讲。  
  
至此，声音就成了一个12行（假设声学特征是12维）、N列的一个矩阵，称之为观察序列，这里N为总帧数。观察序列如下图所示，图中，每一帧都用一个12维的向量表示，色块的颜色深浅表示向量值的大小。  
  
接下来就要介绍怎样把这个矩阵变成文本了。首先要介绍两个概念：

1. 音素：单词的发音由音素构成。对英语，一种常用的音素集是卡内基梅隆大学的一套由39个音素构成的音素集，参见[The CMU Pronouncing Dictionary](http://link.zhihu.com/?target=http%3A//www.speech.cs.cmu.edu/cgi-bin/cmudict)‎。汉语一般直接用全部声母和韵母作为音素集，另外汉语识别还分有调无调，不详述。
2. 状态：这里理解成比音素更细致的语音单位就行啦。通常把一个音素划分成3个状态。

语音识别是怎么工作的呢？实际上一点都不神秘，无非是：  
第一步，把帧识别成状态（难点）。  
第二步，把状态组合成音素。  
第三步，把音素组合成单词。  
  
如下图所示：  
图中，每个小竖条代表一帧，若干帧语音对应一个状态，每三个状态组合成一个音素，若干个音素组合成一个单词。也就是说，只要知道每帧语音对应哪个状态了，语音识别的结果也就出来了。  
  
那每帧音素对应哪个状态呢？有个容易想到的办法，看某帧对应哪个状态的概率最大，那这帧就属于哪个状态。比如下面的示意图，这帧在状态S3上的条件概率最大，因此就猜这帧属于状态S3。  
  
那这些用到的概率从哪里读取呢？有个叫“声学模型”的东西，里面存了一大堆参数，通过这些参数，就可以知道帧和状态对应的概率。获取这一大堆参数的方法叫做“训练”，需要使用巨大数量的语音数据，训练的方法比较繁琐，这里不讲。  
  
但这样做有一个问题：每一帧都会得到一个状态号，最后整个语音就会得到一堆乱七八糟的状态号，相邻两帧间的状态号基本都不相同。假设语音有1000帧，每帧对应1个状态，每3个状态组合成一个音素，那么大概会组合成300个音素，但这段语音其实根本没有这么多音素。如果真这么做，得到的状态号可能根本无法组合成音素。实际上，相邻帧的状态应该大多数都是相同的才合理，因为每帧很短。  
  
解决这个问题的常用方法就是使用隐马尔可夫模型（Hidden Markov Model，HMM）。这东西听起来好像很高深的样子，实际上用起来很简单：  
第一步，构建一个状态网络。  
第二步，从状态网络中寻找与声音最匹配的路径。  
  
这样就把结果限制在预先设定的网络中，避免了刚才说到的问题，当然也带来一个局限，比如你设定的网络里只包含了“今天晴天”和“今天下雨”两个句子的状态路径，那么不管说些什么，识别出的结果必然是这两个句子中的一句。  
  
那如果想识别任意文本呢？把这个网络搭得足够大，包含任意文本的路径就可以了。但这个网络越大，想要达到比较好的识别准确率就越难。所以要根据实际任务的需求，合理选择网络大小和结构。  
  
搭建状态网络，是由单词级网络展开成音素网络，再展开成状态网络。语音识别过程其实就是在状态网络中搜索一条最佳路径，语音对应这条路径的概率最大，这称之为“解码”。路径搜索的算法是一种动态规划剪枝的算法，称之为Viterbi算法，用于寻找全局最优路径。  
  
  
这里所说的累积概率，由三部分构成，分别是：

1. 观察概率：每帧和每个状态对应的概率
2. 转移概率：每个状态转移到自身或转移到下个状态的概率
3. 语言概率：根据语言统计规律得到的概率

其中，前两种概率从声学模型中获取，最后一种概率从语言模型中获取。语言模型是使用大量的文本训练出来的，可以利用某门语言本身的统计规律来帮助提升识别正确率。语言模型很重要，如果不使用语言模型，当状态网络较大时，识别出的结果基本是一团乱麻。  
  
  
这样基本上语音识别过程就完成了。  
  
以上的文字只是想让大家容易理解，并不追求严谨。事实上，HMM的内涵绝不是上面所说的“无非是个状态网络”那么简单，如果希望深入了解，下面给出了几篇阅读材料：  
  
1. Rabiner L R. *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, 1989, 77(2): 257-286.  
**入门必读**。深入浅出地介绍了基于HMM的语音识别的原理，不注重公式的细节推导而是着重阐述公式背后的物理意义。  
  
2. Bilmes J A. *A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models*. International Computer Science Institute, 1998, 4(510): 126.  
详细介绍了用E-M算法训练HMM参数的推导过程，首先讲E-M的基本原理，然后讲解如何应用到GMM的训练，最后讲解如何应用到HMM的训练。  
  
3. Young S, Evermann G, Gales M, et al. *The HTK book (v3.4)*. Cambridge University, 2006.  
HTK Book，开源工具包HTK的文档。虽然现在HTK已经不是最流行的了，但仍然强烈推荐按照书里的第二章流程做一遍，你可以搭建出一个简单的数字串识别系统。  
  
4. Graves A. *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer Berlin Heidelberg, 2012: 15-35.  
基于神经网络的语音识别的入门必读。从神经网络的基本结构、BP算法等介绍到 LSTM、CTC。  
  
5. 余栋, 邓力. *解析深度学习——语音识别实践*, 电子工业出版社, 2016.  
高质量的中文资料非常稀有，推荐买一本。最早把深度学习技术应用于语音识别就是这本书的作者。

[编辑于 2016-07-23](https://www.zhihu.com/question/20398418/answer/18080841" \t "_blank) [98 条评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [禁止转载](https://www.zhihu.com/terms#sec-licence-6)

收起

66赞同反对，不会显示你的姓名

[https://pic1.zhimg.com/da8e974dc_s.jpg](javascript:;)语音识别博士生，剑桥大学工程系

[66 人赞同](https://www.zhihu.com/question/20398418)

楼上张俊博的回答比较仔细的讲解了基础的经典语音识别算法。我想对算法背后的含义做一个简单的解释，对涉及到的特征提取（包括分帧）、音素建模、字典、隐式马尔科夫模型等可以参阅楼上的回答。  
  
语音识别的第一个特点是要识别的语音的内容（比声韵母等）是不定长时序，也就是说，在识别以前你不可能知道当前的 声韵母有多长，这样在构建统计模型输入语音特征的时候无法简单判定到底该输入0.0到0.5秒还是0.2到0.8秒进行识别，同时多数常见的模型都不方便处理维度不确定的输入特征（注意在一次处理的时候，时间长度转化成了当前的特征维度）。一种简单的解决思路是对语音进行分帧，每一帧占有比较短固定的时 长（比如25ms），再假设说这样的一帧既足够长（可以蕴含 足以判断它属于哪个声韵母的信息），又很平稳（方便进行短时傅里叶分析），这样将每一帧转换为一个特征向量，（依次）分别识别它们属于哪个声韵母，就可以 解决问题。识别的结果可以是比如第100到第105帧是声母c，而第106帧到115帧是韵母eng等。 这种思路有点类似微积分 中的『以直代曲』。另外在实际的分帧过程中，还有很多常用技巧，比如相邻两帧之间有所重叠，或引入与临近帧之间的差分作为额外特征，乃至直接堆叠许多语音帧等等，这些都可以让前述的两个假设更可靠。近年来，研究种也出现了一些更新颖的处理方式，比如用.wav文件的采样点取代分帧并处理后的语音帧，但这样的方法在处理速度及性能上 暂时还没有优势。  
  
当我们有了分帧后的语音特征之后，下一步常用的处理是使用某种分类器将之分类成某种跟语音内容相关的类别，如声韵母，这一步通常称作声学模型建模。对于分类目标的选取，最简单的选择可以是词组，或者是组成词组的汉字所对应的音节。但这样的选择方式通常会对训练模型的语音数据提出过高的要求，带来『数据稀疏』的问题，即数据中 很难包含汉语中的所有词组，同时每个词组也很难具有充足的训练样本以保证统计声学模型的可靠性。由于一个词组通常由多个音素的连续发音 构成，常见的音素都包含在国际音标表中，它们具有恰当的数目（通常几十个），以及清晰的定义（由特定的发声器官运动产生），于是音素成了各种语言中的语音识别中都最为常见的 建模选择（汉语的声韵母也是由一到三个音素构成）， 识别中再结合词组到音素的发音字典使用。使用音素也方便对混合语言（如汉语种夹杂英语词汇）进行识别——当然不同母语的人对相同音素的发音也有区别，这是另外一个话题。另外由于人类发生器官运动的连续性，以及某些语言中特定的拼读习惯（比如英语中定冠词『the』在元音和辅音之前有不同读音），会导致发音，尤其是音素的发音受到前后音素的影响，称为『协同发音』。于是 可以进行所谓『上下文相关』的音素（或者考虑到音素实际的拼读，称为音子）分类。比如wo chi le这个序列，可以写为w o ch i l e，这样是普通的『上下文无关』音子序列，也可以考虑前一个音素对当前音素的影响写成sil-w w-o o-ch ch-i i-l l-e （sil表示语音开始前的静音，A-B表示收到A影响的B）或考虑后一个音素的影响写成w+o o+ch ch+i i+l l+e e+sil（sil表示语音结束后的静音，A+B表示受到B影响的A）。实际中以同时考虑前后各一个音素的三音子最为常见，最多也有人使用四音子模型。使用三音子或四音子模型会导致 分类目标的几何增长（如仅仅30个音素就可以扩展出30^3=27000个三音子）， 并再次导致数据稀疏的问题。最常用的解决方法是使用基于决策树的方式对这些三音子或四音子模型进行聚类，对每一类模型进行参数共享以及训练数据的共享。在构建决策树的方式上以及决策树进行自顶向下的 分裂过程中，都可以 导入适当的语音学知识， 将知识与数据驱动的方法进行结合， 同时还可以 减少运算量并在识别中 使用训练数据中未出现的三音子模型等。  
  
有了具体的分类的目标（比如三音子）之后，下面就要选择具体的数学模型进行声学建模。这里可以根据语音学等研究 使用多种线性结构或非线性结构的模型或模型组合。目前最广泛使用的仍然是基于隐式马尔科夫模型的建模方法，即对每个三音子分别建立一个模型，具体可以参见楼上的回答。隐式马尔科夫模型的转移概率密度以几何分布最为常见，但语音合成中也常用高斯分布；观测概率密度函数传统上通常使用 高斯混合模型，也有人使用人工神经网络等，近年来随着深度学习的发展，使用各种深层神经网络的情况 越来越多。最近也有人使用不同方法直接利用递归神经网络进行建模，有一些工作也取得了比较好的效果。但无论使用哪种模型甚至非线性的模型 组合，背后的含义都是假设了对应于每种 类别（三音子）的语音帧在它所对应的高维空间中具有几乎确定的空间分布，可以通过对空间进行划分，并由未知语音帧的空间位置来对语音帧进行正确的分类。  
  
在完成声学模型建模后，就可以基于声学模型对未知语音帧序列进行语音识别了，这一过程通常称为搜索解码过程。解码的原理通常是在给定了根据语法、字典对马尔科夫模型进行连接后的搜索的网络（网络的每个节点可以是一个词组等）后，在所有可能的搜索路径中选择一条或多条最优（通常是最大后验概率）路径（字典中出现词组的词组串）作为识别结果，具体的搜索算法可以有不同的实现方式。这样的搜索可以对时序的语音帧根据其前后帧进行约束；注意使用多状态 隐式马尔科夫模型的理由之一是可以在 搜索中对每个三音子的最短长度施加限制。语音识别任务通常有不同的分类，最困难的问题是所谓大词表连续语音识别，即对可能由数万种日常用词组成的发音自然的语句（比如我们日常随意对话中的语句）进行识别，这样的 问题中通常要 将声学模型同概率语言模型联合使用，即在搜索中导入 统计获得的先验语言层级信息，优点是可以显著的提高识别器的性能，缺点是也会造成识别器明显偏向于识别出语言模型中 出现过的信息。  
  
以上就是我理解的语音识别的原理，包括大致的系统构成和基本设计思路。具体在最前沿的研究和评测 中，通常还需要把许多不同的语音识别器通过各种不同的手段进行系统组合，以便在最终使最终的（组合）系统 能够获得具有互补性的信息，从而得到最佳的识别效果。

[编辑于 2016-07-29](https://www.zhihu.com/question/20398418/answer/71349128" \t "_blank) [10 条评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

6赞同反对，不会显示你的姓名

[https://pic4.zhimg.com/3305f706b_s.jpg](https://www.zhihu.com/people/lifanxi)[**李凡希**](https://www.zhihu.com/people/lifanxi)软件工程师

[6 人赞同](https://www.zhihu.com/question/20398418)

很多知友指出《数学之美》中讨论的只是自然语言处理部分的内容，与语音识别关系不大，对原先这个不太科学的回答可能造成的误导表示抱歉。回答这个问题的本意是分享一些自己所看到过的信息，所以还是把原来的回答保留在下面。  
==  
《数学之美》一书中对自然语言处理的主题有所提及，可以参考一下。我对此是外行，但从这本书中的描述来看，现在自然语言处理技术主要是基于统计学来做的，纯粹的传统意义上的基于规则的语义分析（类似于编译原理中讲的那些）这条路基本上是走不通的。

[编辑于 2014-11-28](https://www.zhihu.com/question/20398418/answer/15016074" \t "_blank) [6 条评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

4赞同反对，不会显示你的姓名

[https://pic1.zhimg.com/07c76dca0_s.jpg](https://www.zhihu.com/people/cancercst)[**刘浩**](https://www.zhihu.com/people/cancercst)IT、语音、产品经理、好奇心、户外、旅游

[4 人赞同](https://www.zhihu.com/question/20398418)

补充个人的理解语音识别，狭义的识别，为文字内容识别；宽泛的识别，可以当作人耳，人能从语音中听出男女、语种、说话人、内容等，则意味着语音识别就是机器耳朵，能有性别、语种方言、声纹（类似指纹概念）、内容识别了。简单而言，目前识别就是2步，训练和测试。拿好理解的语种识别来说：训练，就是提取特征描述语音，是学习过程，例如搞一定量的汉语、英语，训练出2个语种模型；而识别测试，则每一条语音提取特征，然后和这2个语种模型特征对比，给出相似度判断即可。

[发布于 2013-01-05](https://www.zhihu.com/question/20398418/answer/15871744" \t "_blank) [2 条评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

6赞同反对，不会显示你的姓名

[https://pic1.zhimg.com/da8e974dc_s.jpg](https://www.zhihu.com/people/xuh5156)[**徐华**](https://www.zhihu.com/people/xuh5156)程序猿

[6 人赞同](https://www.zhihu.com/question/20398418)

个人认为语音识别可以看成两个步骤的输入法：第一，根据声学模型把声音转换成音节（可以看成拼音），第二，就是就根据统计语言模型把拼音转换成文字了。主要问题是两个模型怎么去训练和使用

[发布于 2013-06-18](https://www.zhihu.com/question/20398418/answer/17564376" \t "_blank) [4 条评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

0赞同反对，不会显示你的姓名

[https://pic1.zhimg.com/da8e974dc_s.jpg](https://www.zhihu.com/people/null122)[**null122**](https://www.zhihu.com/people/null122)

先通过经验模型转换为字符串  
再隐马模式识别转换为自然语言

[发布于 2015-11-17](https://www.zhihu.com/question/20398418/answer/72664099" \t "_blank) [添加评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

4赞同反对，不会显示你的姓名

[https://pic1.zhimg.com/d0dba996a7eead9b3688fe72faafb6d0_s.jpg](https://www.zhihu.com/people/stdlib)[**风清扬**](https://www.zhihu.com/people/stdlib)人工智能/机器学习

[4 人赞同](https://www.zhihu.com/question/20398418)

1. 模拟信号转数字信号，这个太easy啦。设定好采样率，声道数，写一个简单的wav处理程序就可以。  
2. 模型训练：一部分有答案的数据用于训练模型，现在常用的也就是楼上说的HMM模型。训练过程就是提特征+特征训练。特征主要用MFCC等等。  
3. 模型匹配：对测试数据同样提特征，然后进行模型匹配。

[发布于 2012-12-26](https://www.zhihu.com/question/20398418/answer/15816411" \t "_blank) [2 条评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

7赞同反对，不会显示你的姓名

[https://pic4.zhimg.com/eb536b107_s.jpg](https://www.zhihu.com/people/Kubar)[**Kubar**](https://www.zhihu.com/people/Kubar)此人匿名回答的问题比较靠谱，用真名回答…

[7 人赞同](https://www.zhihu.com/question/20398418)

大体的想法还是正确的，不过有一点：  
设计一个“标准”的模拟信号库，与用户输入比对这个，在网络上为你找了一点更详细的介绍：  
  
最早的基于电子计算机的语音识别系统是由AT&T贝尔实验室开发的Audrey语音识别系统，它能够识别10个英文数字。其识别方法是跟踪语音中的共振峰。该系统得到了98%的正确率。到1950年代末，伦敦学院(Colledge of London)的Denes已经将语法概率加入语音识别中。　　1960年代，人工神经网络被引入了语音识别。这一时代的两大突破是线性预测编码Linear Predictive Coding (LPC)， 及动态时间弯折Dynamic Time Warp技术。 语音识别的最重大突破是隐含马尔科夫模型Hidden Markov Model的应用。从Baum提出相关数学推理，经过Labiner等人的研究，李开复最终实现了第一个基于隐马尔科夫模型的大词汇量语音识别系统Sphinx，他也凭借此发明获得1988 美国商业周刊最重要发明奖。此后严格来说语音识别技术并没有脱离HMM框架，可以说现在手机上的语音拨号就脱胎于李开复的语音识别系统。   
  
引用百度百科《语音识别系统》，[http://baike.baidu.com/view/1013816.htm](https://link.zhihu.com/?target=http%3A//baike.baidu.com/view/1013816.htm)

[发布于 2012-08-03](https://www.zhihu.com/question/20398418/answer/15015708" \t "_blank) [添加评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

0赞同反对，不会显示你的姓名

[https://pic1.zhimg.com/da8e974dc_s.jpg](javascript:;)量化投资

是根据统计概率计算出来的，国内做的比较好的是科大讯飞，目前的普通话考试用的就是他们的系统。  
吴军的《数学之美》对语音识别做了个通俗易懂的介绍，有兴趣可以去看看。

[发布于 2015-02-25](https://www.zhihu.com/question/20398418/answer/40545268" \t "_blank) [添加评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

1赞同反对，不会显示你的姓名

[https://pic1.zhimg.com/da8e974dc_s.jpg](javascript:;)知行合一，不亦乐乎

[1 人赞同](https://www.zhihu.com/question/20398418)

门外汉，纯猜测  
  
既然音频能够转换成波形，那么语音识别和图像识别技术在某些方面是共通的。  
因为人说的话都是连贯的，带有语义的，所以语音识别又和全文搜索技术有一定联系。  
  
简单来说  
  
li xiang  
  
这个拼音有很多种可能  
  
如 理想，立项，离乡，莉香等等  
  
但是如果是  
  
li xiang zhe ge ren   
  
那么代表这里的li xiang 是指带的一个人物  
  
再根据语调  
  
可以判断出到第是李湘，还是李想

[发布于 2012-08-03](https://www.zhihu.com/question/20398418/answer/15015991" \t "_blank) [1 条评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

1赞同反对，不会显示你的姓名

[https://pic1.zhimg.com/4e8425304_s.jpg](https://www.zhihu.com/people/waynelu)[**waynelu**](https://www.zhihu.com/people/waynelu)[http://mt.tencent.com](https://link.zhihu.com/?target=http%3A//mt.tencent.com) webapp

[1 人赞同](https://www.zhihu.com/question/20398418)

java 同学可以看看这个：[http://www.cp.jku.at/people/schedl/Research/Development/CoMIRVA/webpage/CoMIRVA.html](https://link.zhihu.com/?target=http%3A//www.cp.jku.at/people/schedl/Research/Development/CoMIRVA/webpage/CoMIRVA.html)

[发布于 2014-07-25](https://www.zhihu.com/question/20398418/answer/28378964" \t "_blank) [添加评论](https://www.zhihu.com/question/20398418) [感谢](https://www.zhihu.com/question/20398418)

[分享](https://www.zhihu.com/question/20398418)

[收藏](https://www.zhihu.com/question/20398418) • [没有帮助](https://www.zhihu.com/question/20398418) • [举报](https://www.zhihu.com/question/20398418) • [作者保留权利](https://www.zhihu.com/terms#sec-licence-1)

**我来回答这个问题**

写回答…

[我要回答](javascript:)

**加入知乎**

与世界分享你的知识、经验和见解

窗体顶端









验证码



已有帐号？[登录](https://www.zhihu.com/question/20398418)

窗体底端

 下载知乎 App

关注问题 2674 人关注该问题

[](https://sugar.zhihu.com/track?vs=1&ai=1137&ut=85dadaf287e140f6b20b841c7ea68d6f&tu=https%3A//zhuanlan.zhihu.com/p/21969681&cg=2&ts=1471072859.64&si=a388b5c82da94772b44007f61f89067f&lu=0&hn=ad-engine.ad-engine.ec22d117&at=click&pf=Windows&az=3&sg=78d8a1ce056fe0db098ed613a3b901fb)https://sugar.zhihu.com/track?vs=1&ai=1137&ut=85dadaf287e140f6b20b841c7ea68d6f&cg=2&ts=1471072859.64&si=a388b5c82da94772b44007f61f89067f&lu=0&hn=ad-engine.ad-engine.ec22d117&pf=Windows&az=3&sg=44ef21d3dae77c8e7159bb5c04b07a19&at=impression

换一换

**相关问题**

* [微软的 Cortana 语音助手说话语调极为逼真像人，这是怎么做到的？](https://www.zhihu.com/question/23283138)17 个回答
* [Nuance 不够好吗，为什么苹果公司要组建自己的语音识别团队？](https://www.zhihu.com/question/21392804)15 个回答
* [移动设备上的离线语音识别需要哪些技术支持，具体如何实现？](https://www.zhihu.com/question/21216684)7 个回答
* [微软的原声翻译在未来会威胁到翻译这职业吗？](https://www.zhihu.com/question/23339363)31 个回答
* [为什么 Deep Learning 最先在语音识别和图像处理领域取得突破？](https://www.zhihu.com/question/21815490)33 个回答
* [刘看山](https://liukanshan.zhihu.com/)

* [移动应用](https://www.zhihu.com/app)

* [加入知乎](https://www.zhihu.com/careers)

* [知乎协议](https://www.zhihu.com/terms)

* [联系我们](https://www.zhihu.com/contact)

© 2016 知乎