●语言学

会话智能代理与语音自动识别

冯志伟① 詹宏伟
(杭州师范大学,杭州 311121)

提 要:本文从会话智能代理的角度,论述语音自动识别的原理和方法,分析特征提取阶段、声学建模阶段和解码阶段的基本原理,最后介绍语音识别研究的历史与现状。

关键词:智能代理;语音自动识别;特征抽取阶段;声学建模阶段;解码阶段
中图分类号:H087 文献标识码:A 文章编号:1000-0100(2018)01-0013-11
DOI 编码:10.16263/j.cnki.23-1071/h.2018.01.003

Conversation Agent and Automatic Speech Recognition

Feng Zhi-wei Zhan Hong-wei
(Hangzhou Normal University, Hangzhou 311121, China)

From the perspective of conversation agent, this paper discusses the theory and approaches of automatic speech recognition, and analyzes the basic principles of feature abstraction stage, acoustic modeling stage and decoding stage. Finally, the past and present of speech automatic recognition are also introduced.

Key words: conversation agent; automatic speech recognition; feature abstraction stage; acoustic modeling stage; decoding stage

1 引言

在自然语言处理(natural language processing)中,人与计算机之间对话与会话的系统称为“会话智能代理”(conversation agent),这个术语中的“会话”也包括“对话”,因为大多数的会话都以对话的方式出现。会话智能代理是一种能够使用自然语言与用户进行交流的计算机程序,它可以帮助用户使用计算机自动地预订机票、回答问题或回复电子邮件。其中的许多问题还与商业会议摘要系统以及其他口语理解系统相关,具有重要的应用价值(冯志伟 2017:183)。

一个完整的会话智能代理系统(conversation agent system)应当包含 6 个组件:语音识别组件(Speech Recognition)、自然语言理解组件(Natural Language Understanding)、自然语言生成组件(Natural Language Generation)、文本—语音合成组件(Text-to-Speech Synthesis)、对话管理组件(Dialogue Manager)和任务管理组件(Task Manager),如图1所示。

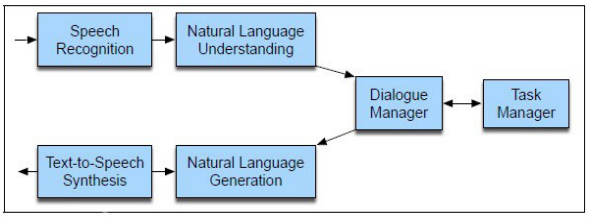


图1 会话智能代理的各组件的简化架构图

在会话智能代理系统的 6 个组件中,根据其功能可以大致分成 3 类:管理组件、输入组件和输出组件。对话管理组件和任务管理组件共同控制会话智能代理系统的整个工作过程。信息操作流程如下:用户的会话输入到语音识别组件后转化为文本形式,自然语言理解组件从中抽取含义,经过管理组件的运算和处理,自然语言生成组件得到文本处理结果,最后文本—语音合成组件将文本处理结果映射到语音,输出对话的最终结果(同上 2014:27)。

系统接受的语音信号来自用户,用户可以通

过电话、PDA(个人数据助理)或笔记本电脑的麦克风向系统发起会话。系统的语音识别组件把这些语音信号转换为单词串,输入会话智能代理系统。语音识别组件是智能会话代理系统的输入端,这个组件在智能会话代理系统中起到举足轻重的关键作用。如果语音识别组件不能有效地识别输入的语音,将会严重地影响到会话智能代理系统的效率。语音识别组件的这种功能叫做语音自动识别(automatic speech recognition,简称 ASR)。语音自动识别的研究目标就是用计算机来建立语音识别系统,把声学信号映射为单词串(Huang et al. 2001)。近年来,语音自动识别的技术在某些限定的领域内已经取得可喜的成果,在会话智能代理系统中已经得到一定程度的应用,这种技术已经逐渐成熟,我们应当密切关注。本文从智能会话代理的角度,尝试以最直观的方式来介绍语音自动识别的方法和技术,以满足读者更新知识的要求。

2　影响语音识别效果的 4 个可变维度

在讨论语音识别的总体结构前,我们先讨论一下影响语音识别效果的 4 个因素,我们把这 4 个因素称为语音识别的"可变维度"。

(1)词汇量的大小:影响语音识别的第一个可变维度是词汇量的大小。如果要识别的话语中不同单词的数量比较小,语音识别就会较容易。只有两个单词的词汇量的语音识别,例如,辨别 yes 还是 no,或者识别只包括 11 个单词的词汇量的数字序列(从英语的 zero 到 nine 再加上 oh),也就是数字识别工作,这种语音识别比较容易。另一方面,对于那些包含 20,000 到 60,000 个单词的大词汇量语音识别,例如,识别人与人之间的电话会话,或者识别广播或电视中的新闻节目,语音识别就相对困难得多。

(2)语音的流畅度和自然度:影响语音识别的第二个可变维度是语音的流畅度、自然度以及是否为对话语音。在孤立单词(isolated word)的识别中,每一个单词被它前后的停顿包围,孤立单词的识别比连续语音的识别容易得多,因为在连续语音的识别中,单词前后彼此连续,必须进行自动切分。连续语音识别这一工作本身的困难程度也各有不同。例如,人对人说话和人对机器说话的流畅度、自然度不同,语音识别的难度也相应不同,前者比后者更难。识别人对机器说话的语音,或者是以阅读语音(read speech)的方式来大声朗读(例如,模拟听写的工作),或者使用语音对话系统来进行转写,都比较容易。在会话智能代理系统中,当人对机器讲话的时候,似乎总是简化自己发出的语音,尽量说得慢一些,说得清楚一些,这样的语音也就比较容易识别。识别两个人以对话语音的方式彼此随意地谈话的语音,例如,转写商业会谈的语音,这样的语音识别就困难得多。

(3)信道和噪声:影响语音识别的第三个可变维度是信道和噪声,它们是信息传递的环境和外部条件。听写以及语音识别的很多实验研究都是在高质量的语音以及头戴扩音器的条件下进行。由于头戴扩音器就可以消除把扩音器放在桌子上时发生的语音失真,因为把扩音器放在桌子上时,说话人的头会动来动去而造成语音失真。任何类型的噪声都会使语音识别的难度加大。因此,在安静的办公室中识别说话人一板一拍的口授比较容易,而识别开着窗子在高速公路上飞驰的汽车中说话人的声音,就困难得多。

(4)说话人的语音特征:影响语音识别的最后一个可变维度是说话人的口音特征和说话人的类别特征。如果说话人说的是标准的语音,或者说话人的语音与系统训练时的数据比较匹配,那么,语音识别就比较容易。反之,语音识别就比较困难。例如,说话人操陌生的口音,或者是儿童的语音(除非语音识别系统是特别地根据这些类型的语音来训练的)。

图$_2$中的数据来自一些最新的语音识别系统,说明在不同的语音识别任务中,误识的单词的大致百分比,这个百分比叫做词错误率(word error rate,简称 WER)。

Task	Vocabulary	Error Rate %
TI Digits	11 (zero–nine, oh)	.5
Wall Street Journal read speech	5,000	3
Wall Street Journal read speech	20,000	3
Broadcast News	64,000+	10
Conversational Telephone Speech (CTS)	64,000+	20

图$_2$　ASR 在不同任务中的词汇量和词错误率

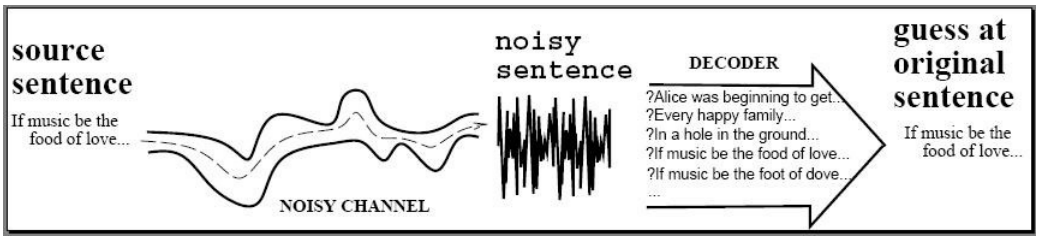
在图$_2$中,广播新闻节目(broadcast news)和电话对话语音(conversational telephone speech,简称 CTS)的错误率是根据特定的训练和测试方案得到的,可以作为一种粗略的估计数字;在这些以不同方式确定的任务中,词错误率的数值变化范围的差别可以达到二倍之多。由于噪声和口音而造成的变化会使错误率增加很多。据报道,对于相同的识别任务,带有浓重日语口音或西班牙语口音的英语的词错误率比母语为英语的人说英语的词错误率约高出 3 至 4 倍。如果汽车噪声的信

噪比(signal-to-noise ratio,简称 SNR) 提高 10 分贝(dB) ,词错误率可能上升 2 至 4 倍。

一般说来,语音识别的词错误率每年都在降低,这是因为语音识别的性能在不断改进中。由于语音识别算法改进和摩尔定律(Moor's law) [②]双重因素的影响,有人估计,在过去的 10 年内,语音识别性能的改进比例大约是每年提高 10% 左右。在本文中,描述的方法应用范围广泛,可以应用于语音识别的各个领域,目前的语音自动识别选择把重点放在大词汇量连续语音识别(large-vocabulary continuous speech recognition,简称 LVCSR) 这个关键性领域的基础性问题。本文中的"大词汇量",指系统包含约二万到六万个单词的词汇; 本文中的"连续",指所有单词是自然地、连续地说出来的。另外,我们将讨论的方法一般是"不依赖于说话人"的(spea-ker-independent) ; 这意味着,这些方法可以识别人的真实语音。由于坚持"大词汇量连续语音识别"这个原则,语音识别取得长足的进展,目前,语音识别系统已经走出实验室,实现实用化和商品化,给现代人的生活和工作带来极大方便,这是在 21 世纪语言学最值得称道的成就。

3　噪声信道模型和隐 Markov 模型

可以从噪声信道模型(noisy channel model) 的角度来看语音识别。这里我们来举例说明噪声信道模型: 源语言的英语句子 if music be the food of love. . . 经过噪声信道变成噪声句子(noisy sentence) ,也就是图$_3$中的声波。为了识别声波,需要对这个噪声句子进行解码,解码时要考虑所有可能的句子,对于每一个句子,都要计算它生成噪声句子的概率,然后,从中选取概率最大的句子,就可以求解出源语言的句子 if music be the food of love. . . 从而达到语音识别的目的。所以,噪声信道模型是一个解码模型。图$_3$具体地说明这个噪声信道模型识别语音的过程。



图$_3$　应用于整个句子语音识别的噪声信道模型

从噪声信道模型的角度来看,语音识别系统的工作就是要搜索一个很大的潜在的源句子空间,并从中选择在生成噪声句子时具有最大概率的句子(冯志伟 2013a) 。建立噪声信道模型需要解决两个问题:

第一是为了挑选出与噪声输入匹配最佳的句子,需要对于最佳匹配有一个完全的度量。因为语音是变化多端的,一个输入句子不可能与这个句子的任何模型都匹配得天衣无缝。因此,要使用概率作为度量,并且说明如何把不同的概率估计结合起来,以便对给定的候选句子的噪声观察序列的概率得到一个全面的估计。第二是因为所有英语句子的集合非常大,需要一个有效的算法使得我们不用搜索所有可能的句子,而只搜索那些有机会与输入匹配的句子。这就是解码问题或搜索问题。抽象地说,语音识别噪声信道模型的总体结构的目标是 "对于给定的某个声学输入 O,在语言 L 的所有句子中,哪一个句子 W 是最可能的句子?"我们可以把声学输入 O 作为单个的符号或"观察"(observation) 的序列来处理。例如,把输入按每 10 微秒切分成音片,每一个音片用它的能量或频度的浮点值来表示。用索引号来表示时间间隔,用有顺序的 o_i表示在时间上前后相续的输入音片。在下面的公式中,用大写字母表示符号的序列,用小写字母表示单个的符号:

$$O = o_1, o_2, o_3, \ldots, o_t$$

类似地,在识别句子时,把句子看成由单词简单地构成的单词串:

$$W = w_1, w_2, w_3, \ldots, w_n$$

不论是声学输入还是句子,上面的这种表示都是简化的假设。在语音识别中,单词通常根据正词法来定义。例如,oak 与 oaks 当作不同的单词来处理; 而助动词 can(can you tell me. . .) 与名词 can(I need a can of. . .) 却当作相同的单词来处理。从隐 Markov 模型(hidden markov model,简称 HMM) 的角度来看,语音识别的任务在于,根据给定的观察 O,求解隐藏在 O 后面的具有最大概率的句子 W. 根据 HMM,对于给定的某个观察 O,具有最大概率的句子 W 可以用每一个句子的两个概率的乘积来计算,并且选乘积最大的句子为所求的句子(冯志

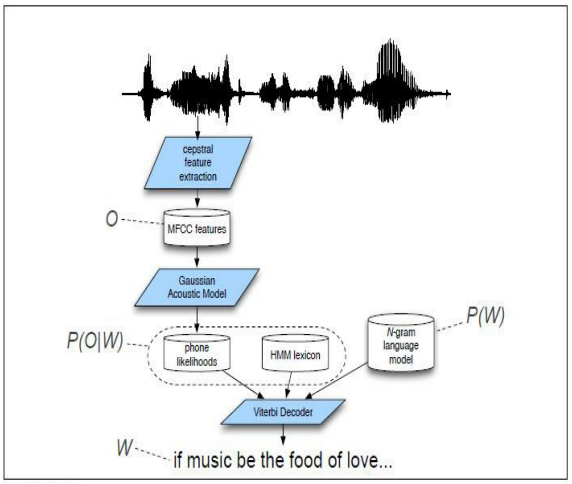
伟 2013b)。HMM 的计算公式如下,其中,P(W) 是先验概率,叫做“语言模型”(language model) ; P(O | W) 是观察似然度,叫做“声学模型”(acoustic model) 。

似然度　先验概率

$$\hat{W} = \arg\max_{W \in L} P(O \mid W)\ \ P(W) \quad ①$$

4　语音识别的 3 个阶段

语音识别可分为 3 个阶段: 特征抽取阶段(feature extraction stage) 、声学建模阶段(acoustic modeling stage) 和解码阶段(decoding stage) ,如图$_4$所示。



图$_4$　语音识别的 3 个阶段

从 HMM 的观点来看,在特征抽取阶段可获取观察值 O,在声学建模阶段可获取观察似然度 P(O | W) 和先验概率 P(W) ,在解码阶段可获取文本 W. 在图$_4$中,输入的是语音,经过这 3 个阶段的处理后,输出的语音识别结果是 if music be the food of love. 在特征抽取阶段,语音的声学波形按照音片的时间框架(通常是 10,15 或 20 毫秒) 来抽样,把音片的时间框架转换成声谱特征(spectral feature) 。每一个时间框架的窗口用矢量来表示,每一个矢量包括约 39 个特征,用以表示声谱的信息以及能量大小和声谱变化的信息。特征信息最普通的表示方法是 Mel 频度倒谱系数(Mel frequency cepstral coefficients,简称 MFCC) 。在图$_4$中,这个阶段具体地用“倒谱特征抽取”(cepstral feature extraction) 表示,抽取到的倒谱特征 MFCC 就是 HMM 中的观察值 O.

在声学建模阶段,对于给定的语言单位(单词、音子和次音子) ,要计算观察到的声谱特征矢量的似然度。例如,我们要使用高斯混合模型(gaussian mixture model,简称 GMM) 分类器,对于 HMM 中与一个音子或一个次音子 W,计算给定音子与给定特征矢量的观察似然度 P(O | W) 。在这个阶段的输出可以用一种简化的方法把它想象成概率矢量的一个序列,在这个序列中,每一个概率矢量对应一个时间框架,而每一个时间框架的每一个矢量是在该时刻生成的声学特征矢量观察 O 与每一个音子单元或次音子单元 W 似然度(phone likelihoods) 。在图$_4$中,这个阶段具体地用“高斯声学模型”(Gaussian acoustic model) 来表示。

在解码阶段,我们取一个声学模型(acoustic model,简称 AM) ,其中包括观察似然度 P(O | W) ,再加上一个 HMM 单词发音词典(HMM lexicon) ,再取一个 N 元语言模型(N-gram language model) ,得到先验概率 P(W) ,把声学模型的观察似然度 P(O | W) 与语言模型的先验概率 P(W) 相结合,输出最可能的单词序列 W. 大多数语音识别系统使用 Viterbi 算法(Viterbi algorithm) 来解码,还采用各种精心设计的提升方法来加快解码的速度,这些方法有剪枝、快速匹配和树结构的词典等。在图$_4$中,这个阶段具体地用“Viterbi 解码”(Viterbi decoder) 来表示。下面我们分别讨论这 3 个阶段。

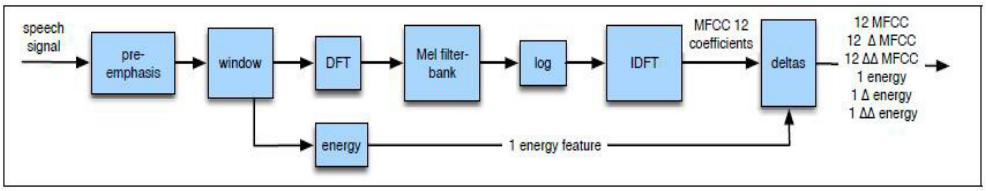
5　特征抽取阶段

我们的目标是描述怎样把输入的波形转换成声学特征矢量(feature vector) 的序列,使得每一个特征矢量代表在一个很小的窗口内的信号的信息。有多种可能的方法来表示这样的信息。迄今最为普通的方法是 Mel 频度倒谱系数 MFCC. MFCC 建立在倒谱(cepstrum) 这个重要的思想基础之上。

首先来讨论模拟语音波形的数字化和量化过程。语音处理的第一步是把模拟信号的表示(首先是空气的压强,其次是扩音器的模拟电信号) 转化为数字信号。这个模拟信号—数字信号转换(analog-to-digital conversion) 的过程分为两步: 第一步是抽样(sampling) ,第二步是量化(quantization) 。信号是通过测定它在特定时刻的幅度来抽样; 每秒钟抽取的样本数叫做抽样率(sampling rate) 。为保证声波测量的精确性,在每一轮抽样中至少需要两个样本: 一个样本用于测量声波的正侧部分,一个样本用于测量声波的负侧部分。每一轮抽样中的样本多于两个时,可以增加抽样幅度的精确性,但是,如果每一轮的样本数目少于

两个，将会导致声波频度的完全遗漏。因此，可能测量的最大频度的波就是那些频度等于抽样率一半的波（因为每一轮抽样须要两个样本）。对于给定抽样率的最大频度叫做 Nyquist 频度（Nyquist frequency）。

对于如像 Switchboard 英语口语语料库这种电话带宽（telephone-bandwidth）的语音来说，8，000赫兹的抽样率已经足够。8，000 赫兹的抽样率要求对于每一秒钟的语音度量 8，000 个幅度，所以有效地存储幅度的度量非常重要。这通常以整数存储，或者是 8 比特，或者是 16 比特。这个把实数值表示为整数的过程叫做量化（quantization），因为这是一个最小的颗粒度（量程规模），所有与这个量程规模接近的值都采用同样的方式来表示。我们把经过数字化和量化的波形记为 x［n］，其中 n 是对于时间的指标。有了波形的数字化和量化的表示，就可以来抽取 MFCC 特征。这个过程可以分为 6 步，如图$_5$所示。



图$_5$　从经过数字化和量化的波形中抽取 39 维的 MFCC 特征矢量序列的过程

从图$_5$可知，语音信号（speech signal）经过预加重（pre-emphasis）、加窗（window）、离散傅里叶变换（DFT）、Mel 滤波器组（Mel filter bank）、对数表示（log）和逆向离散傅里叶变换（iDFT）6 个步骤，得到 12 个 MFCC 系数（MFCC 12 coefficients），与能量特征（energy）一起，成为 Delta 特征（Delta），最后得到 12 个倒谱系数（12 MFCC），12 个 Delta 倒谱系数（12 ΔMFCC），12 个双 Delta 倒谱系数（12 ΔΔMFCC），1 个能量系数（1 energy），1 个 Delta 能量系数（1 Δ energy），1 个双 Delta 能量系数（1ΔΔ energy），一共 39 个 MFCC 特征。我们对图$_5$中的过程进一步加以描述。

（1）预加重

MFCC 特征抽取的第一个阶段是加重高频段的能量，叫做预加重（pre-emphasis）。已经证明，如果观察像元音这样的有浊音的语音片段的声谱，我们会发现，低频端的能量比高频端的能量要高一些。这种频度高而能量下降的现象叫做声谱斜移（spectral tilt），是由于声门脉冲的特性造成的。加重高频端的能量可以使具有较高的共振峰的信息更加适合于声学模型，从而改善音子探测的精确性。这种预加重使用滤波器来进行。

（2）加窗

特征抽取的目的是得到能够帮助我们建立音子或次音子分类器的声谱特征。我们不想从整段的话语或会话中抽取声谱特征，因为在整段的话语或会话中，声谱的变化非常快。从技术上说，语音是非平稳信号（non-stationary signal），因此，语音的统计特性在时间上不是恒定的。所以，我们只想从语音的一个小窗口上抽取声谱特征，从而描述特定的次音子，并大致地假定在这个窗口内的语音信号是平稳的（stationary），也就是假定语音的统计特性在这个区域内是恒定的。为此使用加窗（windowing）的方法，使用窗口来抽取这种大致平稳的语音部分，在窗口内的某个区域内语音信号可不为零或为零，对语音信号运行这个窗口，抽出在这个窗口内的波形。

（3）离散傅里叶变换

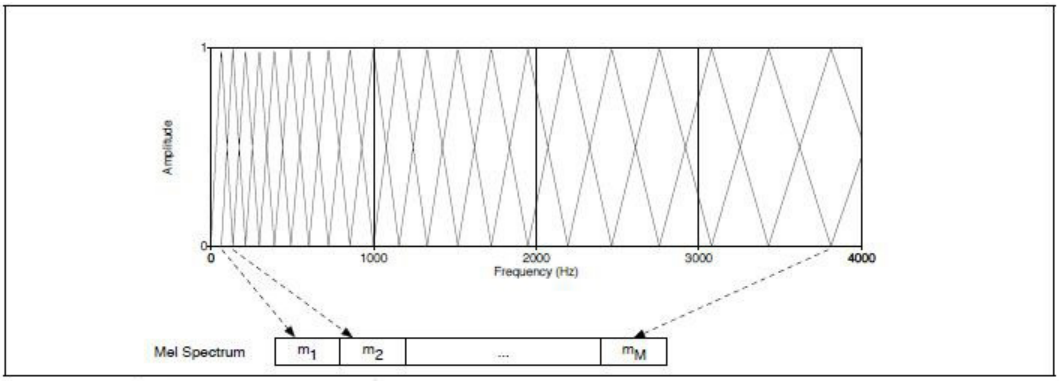
下一步是抽取加窗信号的声谱信息。我们需要知道在不同频带上信号包含的能量有多少。对于抽样的离散时间信号的离散频带，抽取其声谱信息的工具是离散傅里叶变换（discrete Fourier transform，简称 DFT）。计算 DFT 的常用算法是快速傅里叶变换（fast Fourier transform，简称 FFT）。用 FFT 算法来实现 DFT 是很有效的。

（4）Mel 滤波器组

FFT 的计算得到的结果是关于每一个频带上能量大小的信息。然而，人类的听觉并不是在所有的频带上都同样地敏感。它在高频部分（约在一千赫兹）就不太敏感。业已证明，如果在特征抽取时给人类的这种听觉特性建模，就可以改善语音识别的性能。在 MFCC 中使用的这种模型的形式就是把 DFT 输出的频度改变为“美”（Mel）标度，“美”有时也可以直接写为 Mel. 根据定义，如果一对语音在感知上的音高听起来是等距离的，那么，它们就可以用相同数目的“美”分开。在低于 1000 赫兹时，用赫兹表示的频度与“美”标度之间的映射是线性关系，在高于 1000 赫兹时，这种映射是对数关系。“美”的频度 m 可以根据粗糙的声学频度来计算：

② $m = 1127\ln(1+\frac{f}{700})$

计算 MFCC 时，可以建立一个滤波器组(filter bank) 来实现这样的直觉。在这个滤波器组中，收集来自每一个频带的能量，低于 1，000 赫兹的频带的 10 个滤波器遵循线性分布，而其他的高于 1，000 赫兹的频带的滤波器则遵循对数分布。图$_6$说明实现这种思想的三角形滤波器组，图中显示出 Mel 声谱(Mel Spectrum) 。在图$_6$中，每一个三角形滤波器收集来自给定频度范围的能量。



图$_6$　Mel 滤波器组

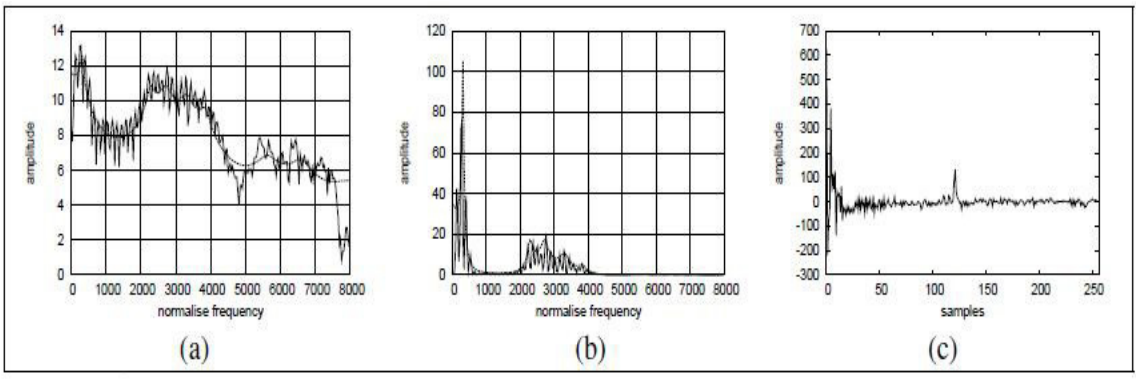
(5) 对数表示

最后，我们使用对数来表示 Mel 声谱的值。在一般情况下，人类对于信号级别的反应是按照对数来计算的; 在振幅高的阶段，人类对于振幅的轻微差别的敏感性比在振幅低的阶段低得多。使用对数来估计特征的时候，对于输入的变化也不太敏感。例如，由于说话人口部运动的收缩或由于使用扩音器等功率变化而导致的输入变化，使用对数来估计时都是不敏感的。

(6) 倒谱的逆向傅里叶变换

使用 Mel 声谱作为语音识别的特征表示是可能的，但是，这样的声谱仍然存在某些问题。MFCC 特征抽取的下一步就是计算倒谱(cepstrum) 。倒谱在语音处理时具有很多长处，它可以改善语音识别的性能。

把声源(source) 和滤波器(filter) 分开，是理解倒谱的一种有效的途径。当带有特定的基本频度的声门的声源波形通过声腔的时候，其形状会带上特定的滤波器特征。但是，声门产生的声源的很多特征(例如，它的基频特征和声门脉冲的细节特征，等等) ，对于区别不同的音子并不是重要的。正是由于这个原因，对于探测音子最有用的信息在于滤波器，也就是声腔的确切位置。如果已知声腔的形状，也就知道将会产生出什么样的音子来。这意味着如果找到一种途径把声源和滤波器区别开来，只给我们提供声腔滤波器，那么，就可以找到音子探测的有用特征。已经证明，倒谱是达到这个目的的一种途径。



图$_7$　振幅表示的声谱(a) ，对数表示的声谱(b) ，倒谱(c)

在图$_7$中，纵轴表示振幅(amplitude) ，声谱的横轴表示规范频度(normalise frequency) ，倒谱的

横轴表示样本(sample) ,为了有助于看清楚声谱,对(a) 和(b) 两个声谱的上部进行平滑处理。为了简单起见,我们忽略 MFCC 中的预加重和 Mel 变形等部分,而只研究倒谱的基本定义。倒谱可以被想象成声谱对数的声谱(spectrum of the log of the spectrum) 。这样的表达似乎有些晦涩。让我们首先来解释比较容易的部分: 声谱对数(log of the spectrum) 。倒谱是从标准的振幅声谱开始的,正如图$_7$(a) 中所示的元音声谱。然后我们对于这个振幅声谱取对数,也就是说,对于振幅声谱中的每一个振幅的值,用它们相应的对数值来表示,如图$_7$(b) 所示。

下一步把这个对数声谱本身也看成似乎是一个波形。换句话说,我们这样考虑图$_7$(b) 中的对数声谱: 把轴上的标记(x 轴上的频度) 去掉,使我们不至于把它想象成声谱,而想象成是正在处理一个正规的语音信号,它的 x 轴表示时间,而不是表示频度。那么,对于这个“假的信号”(pseudo-signal) 的声谱,我们注意到,在这个波中,存在着高频的重复成分: 对于一百二十赫兹左右的频度,小波沿着 x 轴每 1,000 个约重复八次。这个高频成分是由信号的基频引起的,在信号的每一个谐波处,表示为声谱的一个小波峰。此外,在这个“假的信号”中还存在着某些低频成分,例如,对于更低的频度,包络结构或共振峰结构在窗口中有约四个大的波峰。

图$_7$(c) 是倒谱: 倒谱是对数声谱的声谱,上文已经描述过。这个倒谱的英文单词 cepstrum 是由 spectrum(声谱) 前 4 个字母倒过来书写而造出来的,所以叫做“倒谱”。图中的倒谱在 x 轴上的标记是样本。这是因为倒谱是对数声谱的声谱,我们不再理会声谱的频度领域,而回到时间领域。已经证明,倒谱的正确单位是样本。

细心地检查这个倒谱会看到,120 附近有一个大的波峰,相当于 F0,表示声门的脉冲。在 x 轴的低值部分,还存在着其他的各种成分。它们表示声腔滤波器(舌头的位置以及其他发音器官的位置) 。因此,如果对于探测音子有兴趣,那么,就可以使用这些比较低的倒谱值。如果对于探测音高有兴趣,那么,就可以使用较高的倒谱值。为抽取 WFCC,一般只取头 12 个倒谱值。这 12 个参数仅仅表示关于声腔滤波器的信息,它们与关于声门声源的信息的区别是泾渭分明的。

已经证明,倒谱系数有一个非常有用的性质: 不同的倒谱系数之间的方差(variance) 倾向于不相关。而这对于声谱是不成立的,因为不同频带上的声谱系数是相关的。倒谱特征不相关这个事实意味着,高斯声学模型或高斯混合模型不必表示各个 MFCC 特征之间的协方差(covariance) ,这就大大地降低参数的数目。倒谱还可以更加形式化地定义为信号的 DFT 的对数振幅的逆向 DFT,也就是 iDFT.

(7) Delta 特征与能量

从前面的介绍可知,在用逆 DFT 收取倒谱时,每一个帧有 12 个倒谱系数。下面我们再加上第 13 个特征: 帧的能量。能量与音子的识别是相关的,因此,它是探测音子的一个有用的线索(元音和咝音比塞音具有更多的能量) 。一个帧的能量是该帧在某一时段内的样本幂的总和,因此,从时间样本 t_1 到时间样本 t_2 的窗口内,信号 x 的能量是

$$③ \quad Energy = \sum_{t=t_1}^{t_2} x^2[t]$$

语音信号的另外一个重要的事实是: 从一个帧到另一个帧,语音信号是不恒定的。共振峰在转换时的斜坡的变化,塞音从成阻到爆破的变化,这些都可能给语音的探测提供有用的线索。由于这样的原因,我们还可以加上倒谱特征中与时间变化有联系的一些特征。

我们使用对于 13 个特征每一个特征都加上 Delta 特征(Delta feature) 或速度特征(velocity feature) 以及加上双 Delta 特征(double Delta feature) 或加速度特征(acceleration feature) 的办法来做到这一点。这 13 个 Delta 特征中的每一个特征表示在相应的倒谱/能量特征中帧与帧之间的变化,而这 13 个双 Delta 特征中的每一个特征表示在相应的 Delta 特征中帧与帧之间的变化。在给 12 个倒谱特征加上能量特征并进一步添加 Delta 特征和双 Delta 特征之后,我们最后得到如下 39 个 MFCC 特征:

表$_1$　39 个 MFCC 特征

12 个倒谱系数
12 个 Delta 倒谱系数
12 个双 Delta 倒谱系数
1 个能量系数
1 个 Delta 能量系数
1 个双 Delta 能量系数

关于 MFCC 特征的最有用的事实之一就是倒谱系数倾向于不相关。这一事实使得声学模型变得更加简单。上文介绍过语音识别中的特征提取

阶段，说明怎样从波形抽取表示声谱信息的 MFCC 特征，并在每 10 毫秒内产生 39 个 MFCC 特征矢量。

6　声学建模阶段

现在来介绍语音识别的声学建模阶段，研究怎样计算这些特征矢量与给定的 HMM 状态的似然度。这个输出的似然度是通过 HMM 的概率函数 B 来计算的。概率函数 B 也就是观察似然度的集合。对于给定的单独状态 q_i 和观察 o_t，在矩阵 B 中的观察似然度是 $p(o_t \mid q_i)$，我们把它叫做 $b_t(i)$。在词类标注中，每一个观察 o_t 是一个离散符号（一个单词），我们只要数一数在训练集中某个给定的词类标记生成某个给定的观察的次数，就可以计算出一个给定的词类标记生成一个给定观察的似然度。不过，在语音识别中，MFCC 矢量是一个实数值，不可能使用计算每一个这样的矢量出现的次数的方法来计算给定的状态（音子）生成 MFCC 矢量的似然度，因为每一个矢量都有自己的独特性，它们各不相同。

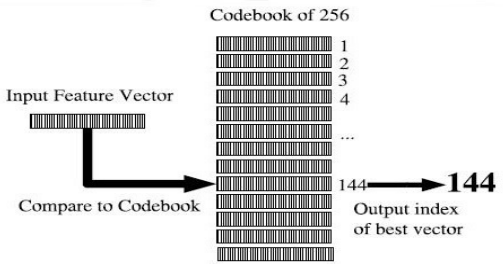
不论在解码时还是在训练时，都需要一个能够对于实数值的观察 o_t 计算 $p(o_t \mid q_i)$ 的观察似然度函数。在解码时，我们有一个观察 o_t，需要对于每一个可能的 HMM 状态，计算概率 $p(o_t \mid q_i)$，使得我们能够选择出最佳的状态序列。为此需要进行矢量的量化。有一个办法可以使 MFCC 矢量看起来像可以记数的符号，这个办法就是建立一个映射函数把每一个输入矢量映射为少量符号中的一个符号。然后就可以使用计算这些符号的方法来计算概率。这种把输入矢量映射为可以量化的离散符号的思想，叫做矢量量化（vector quantization，简称 VQ）。虽然矢量量化做起来就像现代的 LVCSR 系统中的声学模型那样简单，但是，这是一个行之有效的步骤在 ASR 各色各样的领域中起着重要作用，所以，我们使用矢量量化作为讨论声学模型的开始。

在矢量量化时，我们通过把每一个训练特征矢量映射为一个小的类别数目的方法，建立起一个规模很小的符号集，然后，分别使用离散符号来表示每一个类别。更加形式地说，一个矢量量化系统是使用 3 个特征来刻画的，这 3 个特征分别是码本（codebook）、聚类算法（clustering algorithm）和距离测度（distance metric）。

码本是可能类别的表，是组成词汇 $V = \{v_1, v_2, \ldots, v_n\}$ 的符号的集合。对于码本中的每一个代码 v_k，要列出模型矢量（prototype vector），叫做码字（vector word）。码字是一个特定的特征矢量。例如，如果选择使用 256 个码字，就可以使用从 0 到 255 的数值来表示每一个矢量。由于我们使用一个 8 比特的数值来表示每一个矢量，所以叫做 8 比特的矢量量化（8-bit VQ）。这 256 个数值中的每一个数值都与一个模型化的特征矢量相关联。

我们使用聚类算法来建立码本，聚类算法把训练集中所有的特征矢量聚类为 256 个类别。然后，从这个聚类中选择一个有代表性的特征矢量，并把它作为这个聚类的模型矢量或码字，经常使用 K-均值聚类（K-means clustering）。

一旦建立这样的码本，就可以把输入的特征矢量与 256 个模型矢量相比较，使用某种距离测度来选择最接近的模型矢量，用这个模型矢量的索引来替换输入矢量，这个过程如图$_8$所示。



图$_8$　矢量量化（VQ）过程

从图$_8$可以看出，在矢量量化时，把输入的特征矢量（input feature vector）与码本中的每一个码字相比较，使用某种距离测度选择出最接近的条目，输出最接近的码字的索引（output index of best vector）。矢量量化 VQ 的长处在于，由于类别的数目有限，当使用状态来标注和归一化的时候，对于每一个类别 v_k，通过简单地数一数该类别在某一个训练语料库中出现次数的方法，就可以计算出给定的 HMM 状态或次音子生成该类别的概率。

聚类过程和解码过程都要求进行距离测度或失真测度（distortion metric）的计算，以便说明两个声学特征矢量的相似程度。距离测度用于建立聚类，找出每一个聚类的模型矢量，并对输入矢量与模型矢量进行比较。声学特征矢量的最简单的距离测度是 Euclidean 距离（Euclidean distance）。Euclidean 距离是在 N 维空间中由两个矢量定义的两个点之间距离。在实际应用中，我们使用“Euclidean 距离”这个短语经常意味着“Euclidean 距离的平方”。Mahalanobis 距离（Mahalanobis distance）是一个稍微复杂的距离测度，这样的距离测度要考虑到每一个维度中不同的方差。

当给一个语音信号解码时，为了使用矢量量

化来计算对于给定的 HMM 状态 q_j 特征矢量 o_t 的声学似然度，要计算 N 个码字中的每一个码字的特征矢量之间的 Euclidean 距离或 Mahalanobis 距离，选择最接近的码字，得到码字索引 v_k. 然后，先计算 HMM 定义的似然度矩阵 B，找出对于给定 HMM 的状态 j，码字索引 v_k 的似然度:

$$④ \quad \hat{b}_j(o_t) = b_j(v_k)$$

其中，v_k 是最接近矢量 o_t 的码字。

矢量量化的优点是计算起来非常容易，而且只需要很小的存储。尽管有这样的优点，矢量量化还不是语音处理的一个好模型。因为在矢量量化中数量很小的码字不足以捕捉变化多端的语音信号。而且，语音现象并不简单地是一个范畴化的、符号化的过程。因此，现代语音识别算法一般不使用矢量量化来计算声学似然度，而是直接根据实数值的、连续的输入特征矢量来计算观察概率。这些声学模型是建立在连续空间上计算概率密度函数(probability density function，简称 pdf) 的基础之上的。目前最常用的计算声学似然度的方法是高斯混合模型的概率密度函数(pdf) ，此外，还可使用神经网络(neural network) ，支持向量机(support vector machines，简称 SVMs) 和条件随机场(condition random fields，简称 CRFs) 等方法。

7　解码阶段

语音识别的最后一个阶段是解码阶段(decoding stage) ，在解码阶段，我们取一个声学模型，其中包括声学似然度的序列，再加上一个 HMM 的单词发音词典，再取一个语言模型(language model，简称 LM，一般是一个 N 元语法) ，把声学模型与语言模型结合起来，采用 Viterbi 算法进行解码，发现对于给定声学事件具有最大概率的单词序列，得到语音识别的结果。

图9 是语音识别系统总体结构图，图中显示从语音波形经过特征抽取、声学建模和解码等阶段，最后输出英语的单词串 I need a. . . 的过程。

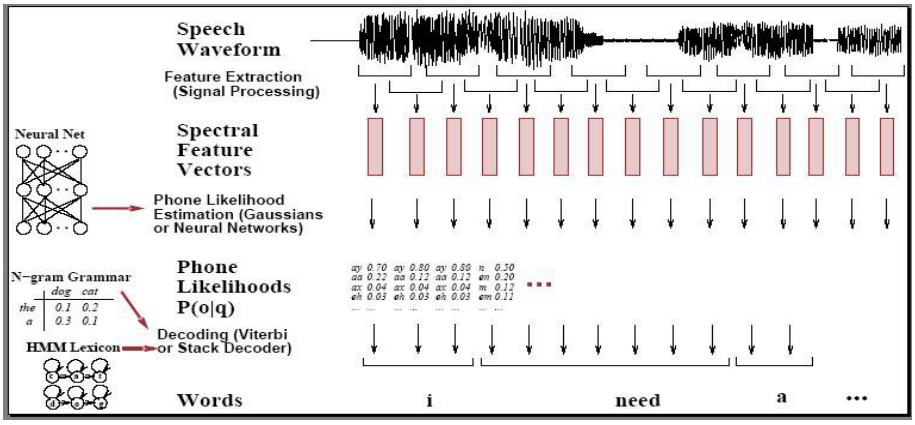


图9　语音识别系统总体结构图

近年来，语音自动识别的研究发展迅速，已经进行商品化的开发，在人机对话、口语机器翻译、智能人机接口和会话智能代理等领域中得到广泛应用。在会话智能代理系统中，语音识别组件接受音频信号，经过语音识别之后返回一个与音频信号相应的单词串进行输入。

会话智能代理中的语音识别组件在许多方面都可能做特殊优化。例如，用于听写或转录的基于大词汇量的语音识别组件，专注于识别任意话题的使用任意单词的任意句子。但是，对于具有领域特殊性的对话系统，识别任意的、各色各样的句子并没有多大作用。这时，语音识别组件需要识别的句子仅仅是那些可以被自然语言理解组件所理解的句子。因此，商业的对话系统普遍使用基于有限状态文法(finite state grammar) 的非概率语言模型，这些语言模型的文法通常都由人工编写，并明确指定系统能理解的所有回答，并不要求能够识别任意话题的使用任意单词的任意句子。

在会话智能代理系统中，因为用户对系统所说的话与系统所处的对话状态密切相关，系统所用的语言模型通常是依赖对话状态的。例如，如果会话智能代理系统刚刚问用户 What city are you departing from? (你要从哪个城市出发?) ，语音识别组件的语言模型会被约束为只包括城市名，或者可能只有形如 I want to (leave | depart) from [CITYNAME] (我想从 [CITYNAME] (出发 |

离开)) 的句子。这种特定于对话状态的语言模型一般由人工编写的有限状态语法或上下文无关语法(context-free grammar)组成,每个对话状态对应一个特定的语法。在某些会话智能代理系统中,如果要识别的句子的集合很大,就可以使用一个 N-元语言模型(N-gram language model)来取代有限状态文法,这个语言模型的概率近似为在对话状态中的条件概率。

无论是使用有限状态文法、上下文无关文法,还是使用 N-元语言模型,我们将这些依赖于对话状态的语言模型称为约束文法(restrictive grammar)。当会话智能代理系统希望约束用户对系统的上一个话段做出回应时,就可以使用约束文法;当系统希望用户有多个选择时,就可以将这个特定状态的语言模型与一个更普遍的语言模型融合在一起即可,选择策略可以根据用户被授予的主动权进行调整(冯志伟 余卫华 2015:80)。对话以及听写等其它应用中的语音识别有一个特征,即说话人的标识在许多话段中都保持不变,这时可以使用说话人自适应技术来改进语音识别的性能。

8　语音自动识别的历史与现状

语音自动识别的早期,只能识别单个的语音。例如,20 世纪 20 年代出现的世界上第一台能够识别语音的机器,是一个名字叫做 Radio Rex 的商品玩具狗。当人们说 Rex 的时候,狗就会在人的叫唤声的控制下走过来。在 20 世纪 40 年代末 50 年代初,Bell 实验室的系统可以识别一个单独说话人的 10 个数字中的任何一个,识别正确率达到 97% 至 99%。Fry 和 Denes 在伦敦大学院建立一个音位识别系统,根据类似的模式识别原则,该系统能够识别英语中的 4 个元音和 9 个辅音。Fry 和 Denes 的系统首次使用音位转移概率来对语音识别系统进行约束(Huang et al. 2001)。

在 20 世纪 60 年代末 70 年代初,语音自动识别技术有较大的突破:(1)出现一系列的特征抽取算法,在语音中的应用倒谱处理,在语音编码中研制线性预测编码(linear predictive coding,简称 LPC)。(2)提出一些处理翘曲变形(warping)的方法,在与存储模式匹配时,通过展宽和收缩输入信号的方法来处理说话速率和切分长度的差异。解决这些问题的最自然的方法是动态规划,在研究这个问题的时候,同样的算法被多次地重新提出。Itakura 把动态规划的思想和 LPC 系数相结合并首先在语音编码中使用。他建立的系统可以抽取输入单词中的 LPC 特征并使用动态规划的方法把这些特征与所存储的 LPC 模板相匹配。这种动态规划方法的非概率应用是对输入语音进行模板匹配,叫做动态时间翘曲变形(dynamic time warping)。(3)HMM 的兴起。1972 年前后,有两个实验室独立地应用 HMM 来研究语音问题。一方面的应用是由一些统计学家的工作引起的,Baum 和他的同事们在普林斯顿(Princeton)的国防分析研究所研究 HMM,并把它应用于解决各种预测问题。James Baker 在卡耐基梅隆大学(Carnegie-Mellon University)做研究生期间,学习 Baum 等人的工作,并把他们的算法应用于语音处理。与此同时,在 IBM 华生研究中心,Jelinek,Mercer 和 Bahl 独立地把 HMM 应用于语音研究,他们在信息论模型方面的研究受到香农(Shannon)的影响。IBM 的系统和 Baker 的系统非常相似,特别是他们都使用贝叶斯方法。他们之间早期工作的一个不同之处是解码算法。Baker 的 DRAGON 系统使用 Viterbi 动态规划解码,而 IBM 系统则应用 Jelinek 的栈解码算法。Baker 在建立语音识别公司的 DRAGON 系统之前,曾经短期参加过 IBM 小组的工作。IBM 的语音识别方法在 20 世纪末期完全地支配这个领域。IBM 实验室确实是把统计模型应用于自然语言处理的推动力量,他们研制基于类别的 N 元语法模型,研制基于 HMM 的词类标注系统,研制统计机器翻译系统,他们还使用熵和困惑度作为评测的度量(Huang et al. 2001)。

此后,HMM 逐渐在语音处理界流传开来。这种流传的原因之一是由于美国国防部高级研究计划署(the Advanced Research Projects Agency,简称 ARPA)发起一系列的研究和开发计划。第一个 5 年计划开始于 1971 年。第一个 5 年计划的目标是建立基于少数说话人的语音识别系统,这个系统使用一个约束性的语法和一个词表(包括 1,000 单词),要求语义错误率低于 10%。ARPA 资助 4 个系统,而且对它们进行比较。这 4 个系统是:系统开发公司的系统(System Development Corporation,简称 SDC);Bolt,Beranek 和 Newman(BBN)的 HWIM 系统;卡耐基梅隆大学的 Hearsay-II 系统;Harpy 系统。其中,Harpy 系统使用 Baker 的基于 HMM 的 DRAGON 系统的一个简化版本,在评测系统时得到最佳的成绩。对于一般的任务,这个系统的语义正确率达到 94%,这是唯一达到 ARPA 计划原定目标的系统。

从 20 世纪 80 年代中期开始,ARPA 资助一些新的语音研究计划。第一个计划的任务是“资

源管理”(Resource Management,简称 RM)。这个计划的任务与 ARPA 早期的课题一样,主要是阅读语音(说话人阅读的句子的词汇量有 1,000 个单词)的转写(也就是语音识别),但这个系统还包括一个不依赖于说话人的语音识别装置。其他的任务包括华尔街杂志(*Wall Street Journal*)的句子阅读识别系统,这个系统开始时的词汇量限制在 5,000 个单词之内,最后的系统已经没有词汇量的限制。事实上,大多数系统已经可以使用约六万个单词的词汇量。后来的语音识别系统识别的语音已经不再是阅读的语音,而是可以识别更加自然的语音。其中识别广播新闻的系统可以转写广播新闻,包括转写那些非常复杂的广播新闻。例如,街头现场采访的新闻;还有 CallHome 系统、CallFriend 系统和 Fisher 系统可以识别朋友之间或者陌生人之间在电话里的自然对话。空中交通信息系统(air traffic information system,简称 ATIS)是一个语音理解的课题,它可以帮助用户预定飞机票,回答用户关于可能乘坐的航班、飞行时间、日期等方面的问题。

ARPA 课题大约每年进行一次汇报,参加汇报的课题除了 ARPA 资助的课题之外,还有来自北美和欧洲的其他“志愿者”系统,在汇报时,彼此测试系统的单词错误率和语义错误率。在早期的测试中,那些赢利的公司一般都不参加比赛,但是,后来很多公司开始参与。ARPA 比赛的结果,促进各个实验室之间广泛地彼此借鉴和交流技术,因为在比赛中很容易看出,在过去一年的研究里,什么样的思想有助于减少错误,而这后来大概就成为 HMM 模型传播到每一个语音识别实验室的重要因素。ARPA 的计划也造就很多有用的数据库,这些数据库原来都是为了评估而设计的训练系统和测试系统(如 TIMIT,RM,WSJ,ATIS,BN,CallHome,Switchboard,Fisher),但是,后来都在各个总体性的研究中得到使用(Jurafsky et al. 2009)。

我国的语音自动处理起步稍晚,但实力已经走在国际前沿。我国语音自动处理领域最突出的代表是科大讯飞。科大讯飞在智能语音技术领域有着长期的研究积累,并在中文语音合成、语音识别和口语评测等多项技术上拥有国际领先的成果。由于科大讯飞拥有自主知识产权的世界领先智能语音技术,他们已推出从大型电信级应用到小型嵌入式应用,从电信、金融等行业到企业和家庭用户,从 PC 到手机到 MP3/MP4/PMP 和玩具,能够满足不同应用环境的多种产品。科大讯飞占有中文语音技术市场 60% 以上市场份额,语音合成产品市场份额达到 70% 以上,开发伙伴超过 500 家,以科大讯飞为核心的中文语音产业链已初具规模(冯志伟 2013a)。

人工智能的重要环节是人机语音交互,其目标是使人与机器之间沟通变得像人与人沟通一样简单。在人工智能的研究中,语音自动处理技术尤为重要。让机器说话,用的是语音合成技术;让机器听懂人说话,用的是语音识别技术。因此,语音自动处理技术的应用前景非常广阔(教育部语言文字信息管理司 2009)。

注释

①冯志伟是杭州师范大学“钱塘学者”讲座教授。

②当价格不变时,集成电路上可容纳的晶体管数目,约每隔 18 个月增加 1 倍,性能也将提升 1 倍。

参考文献

冯志伟. 语言学正面临战略转移的重要时刻[J]. 南开语言学刊,2013a(1).

冯志伟. 隐马尔可夫模型及其在自动词类标注中的应用[J]. 燕山大学学报,2013b(3).

冯志伟. 言语行为理论和会话智能代理[J]. 外国语,2014(1).

冯志伟　余卫华. 智能会话代理系统中的 BDI 模型[J]. 外国语,2015(2).

冯志伟. 自然语言计算机形式分析的理论与方法[M]. 北京:中国科学技术大学出版社,2017.

教育部语言文字信息管理司. 文语转换与语音识别系统语言文字评测规范(草案)[A]. 中国语言生活绿皮书[C]. 北京:语文出版社,2009.

Jurafsky, D., Martin, J. *Speech and Language Processing*[M]. Upper Saddle River: Prentice Hall, 2009.

Huang, X.-D., Acerd, A., Hon, H.-W. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*[M]. Upper Saddle River: Prentice Hall, 2001.

定稿日期:2017-11-24　　【责任编辑　谢　群】