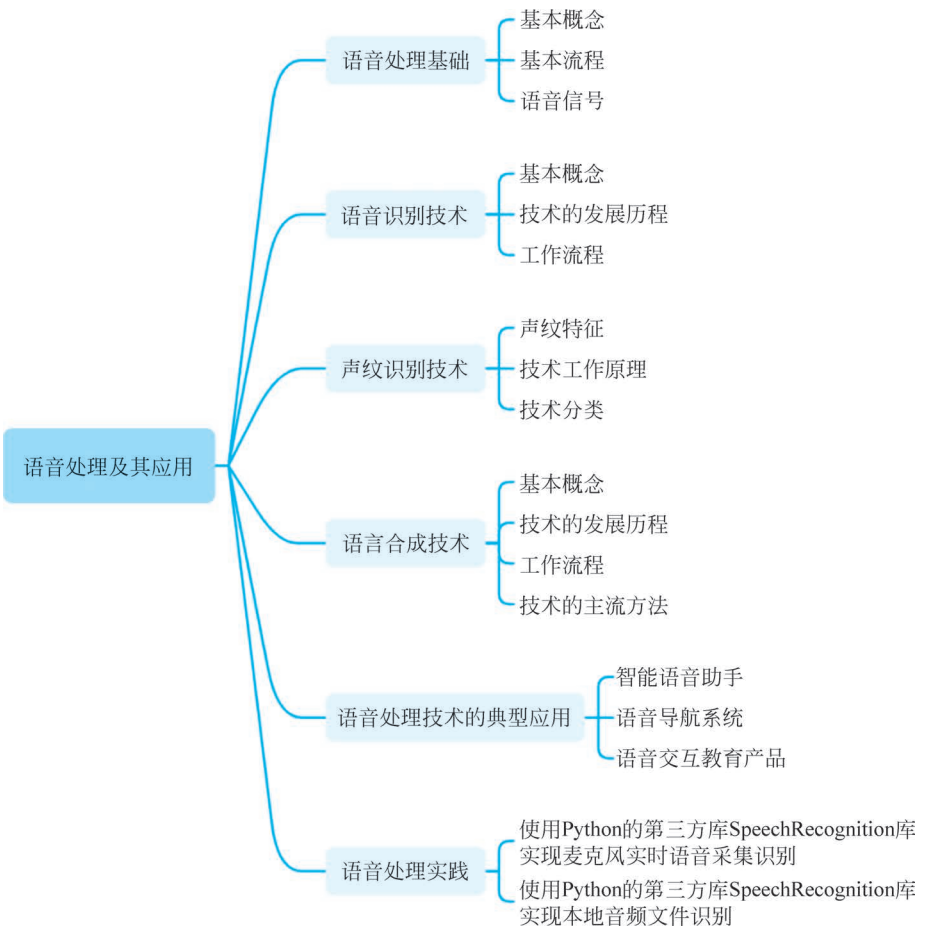
第 3 章 语音处理及其应用

思维导图：



学习目标：

- 了解语音处理的基本概念，语音信号的特征和语音信号处理的基本流程
- 了解语音识别、声纹识别、语音合成等核心技术的概念以及发展历程
- 理解语音识别和语音合成技术的基本工作流程，熟悉语音合成的主流方法
- 理解语音处理技术在智能助手、语音车载系统、语音交互教育产品等典型领域的应

用场景及技术实现路径

- 能够使用 Python 的第三方库 SpeechRecognition 自主编程实现简单的语音识别和本地音频文件识别

3.1 语音处理基础

随着深度学习、云计算、大数据等新一代信息技术的快速发展，人类正进入人工智能与万物互联时代，人类和机器的交互正逐渐进入以“语音交互为主、键盘触摸手势为辅”的阶段。作为万物互联时代最重要、最直接的入口，中国的智能语音技术研发及其产业应用推广近年来取得了长足进步，并长期占据国际领先地位。“能理解会思考”的智能语音技术是指能使信息时代的各种机器像人一样具有“能听会说”和“能理解会思考”的感知和认知能力的智能技术，是人工智能技术产业的核心发展领域。

2020 年，中国产业研究报告网发布的《2021—2027 年中国语音识别系统行业深度研究与发展前景预测报告》指出，目前语音识别有两个方向：一是对词汇量比较大的连续语音进行语音识别；二是发展小型方向，研究适用于便携式产品使用的语音识别。词汇量比较大的语音识别主要应用于大型计算机系统，或者是和互联网结合的语音服务系统。发展小型方向主要是生产带有语音识别性能的语音识别芯片，两者各有各的前景。

3.1.1 语音处理概述

1. 语音处理基本概念

语音是人类传递信息的一种最主要、最有效、最方便的交流形式。语言是人类特有的交流方式，而声音又是人类比较常用的交流工具，是传递信息的主要手段，所以，语音信号是人们情感交流以及思想沟通的主要途径。语音处理作为一门综合性学科，致力于研究语音发声机制、语音信号的统计特性、自动语音识别、机器语音合成以及语音感知等多种处理技术。其中，语音合成与语音识别是人工智能领域的两个重要技术，它们共同构成了人机交互的核心组成部分。语音合成可以将文本转换为人类听觉系统能够理解和接受的声音，从而实现与计算机或其他设备的交互。语音识别则可以将人类的语音信号转换为文本，实现人机的双向沟通。

2. 语音信号

语音信号是人与人交流的自然媒介，它包含丰富的信息，如语义、情感和身份特征。在语音处理的整个流程里，从最初的文本输入到最终语音输出，语音信号贯穿始终。

1）语音信号的产生

在物理学中，声音是由物体的振动产生的，正在发声的物体称为声源。物体在 1 秒钟内振动的次数称为频率，单位是赫兹（Hz），人的耳朵可以听到 20～20 000Hz 的声音，其中最敏感是 1000～3000Hz 的声音。人类语音信号的产生和感知是一个极其复杂的过程，主要可以分为三个阶段：语音的产生、语音的传递和语音的感知。

（1）语音的产生。

人类的发音器官包括肺、气管、喉、咽、鼻、口，它们共同组成一套复杂的发音系统。其中喉的部分称为声门，而从声门到嘴唇的呼气通道称为声道。声道是语音产生以后在人体内

传播的通道。语音的产生有两种不同的方式，分别为声带的震动和声道窄部产生的涡流。声道好比一个滤波器或者共鸣系统，当声音经过声道时，频谱将会发生改变，同时口唇和鼻腔也会使声音的频谱发生改变，声源的不同以及声道形状、嘴型等都会影响声音的音位。每个人的发音器官存在很大的差异性，这也导致了不同人的语音存在一定的差异性。

（2）语音的传递。

语音以声波的形式通过空气等媒介传播。在这个过程中，语音会受到来自环境因素的干扰，例如噪音信号等会造成语音的失真。

（3）语音的感知。

语音的感知是由人耳和大脑共同构成的复杂的听觉系统来完成的。首先，人耳接收到语音的声波信号后会转化成电信号，传递给大脑皮层，然后由大脑内复杂的听觉神经元进行感知。大脑可以感知的语音信号包含音高、音强、音长、音色和语调等复杂信息，从而听话者能准确地判断说话人的意思。

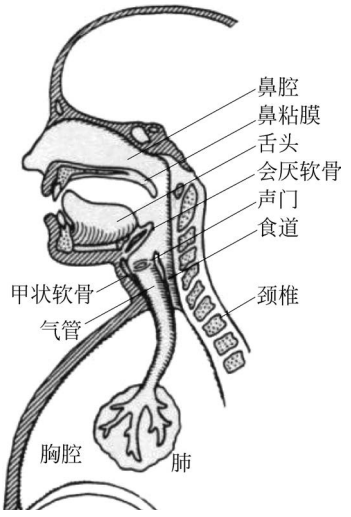


图 3.1　语音发音系统

2）语音的声学特征

语音的声学特征是指各种语音音频信号在声学上的特征。这些特征是通过语音信号的产生、传输和接收过程中的声学效应产生的。因此，在理解语音的声学特征时，需要考虑语音的基本单位——音素（Phoneme），以及声学参数，如频率（Frequency）、振幅（Amplitude）、时长（Time）、共振（Resonance），等等。

语音信号是由一系列较小的语音单元构成的。这些单元被称为音素。音素是语音的最小基本单位。它们被用来构建单词、短语和句子。音素有元音和辅音两种类型。元音由良好的声音质量和长短程度特征定义，辅音由有息音、无息音和破裂音组成。

语音以声波的方式在空气中传播。声波是一种纵波，它的振动方向和传播方向是一致的。声波有一些物理意义上的描述，而从语音学角度看，它具有一些其他特征。声波从声源向四面八方传播，它的频率指在单位时间内声波的周期数。而波长（Wave Length）指声波中两个波峰之间相隔的时间距离。波长是用声波的传播速度/声波的频率。频率越高，波长越短；频率越低，波长越长。从物理描述上看，声波具有两个参数：一个是频率；另一个是振幅。声波的频率是指一个声音波形中每秒的振荡周期数，单位是赫兹（Hz），声音的频率与声音的音高有关。声音的频率高，声音就高；声音的频率低，声音就低。振幅是声波在传播过程中能量的大小，单位是分贝（dB）。在语音中，振幅通常用来表征语音的响度和音量。在荒郊野外大声呼喊，必然振幅大，响度大；在近处低声交头接耳，必然振幅小，响度小。

时长是声音的持续时间，单位为秒（s）。在语音中，时长通常用于描述元音的持续时间和辅音的持续时间。音素的声学特征与其时长有关。例如，元音的声学特征被定义为其始音、高峰和次谷之间的时长；辅音的声学特征则被定义为其始音和尾音之间的时长。

共振是声波在特定频率下放大或减弱的形式，单位是 dB。在语音中，共振通常用来描述元音的音高和声音的质量。元音的声音质量与其所包含的共振特征成正比，而辅音的声音质量则取决于其所在音素的元音共振特征。

总体而言，语音的声学特征是指各种语音音频信号在声学上的特征。这些特征对于语音的理解和识别非常重要，因此对于不同语种的学习和研究都具有重要意义。通过对语音声学特征的深入了解，我们能够更好地理解语音的产生与传递，也能更好地进行语音信号的分析和处理。

3. 语音信号处理的基本流程

语音处理的核心在于理解语音信号的物理特性（频率、振幅、频谱），并利用数字化技术（采样、量化、编码）将其转换为计算机可处理的形式。语音信号处理的基本流程涵盖语音信号采集与预处理、特征提取、模式识别等环节，各环节协同实现语音识别和语音合成等应用，具体的处理流程如图 3.2 所示。

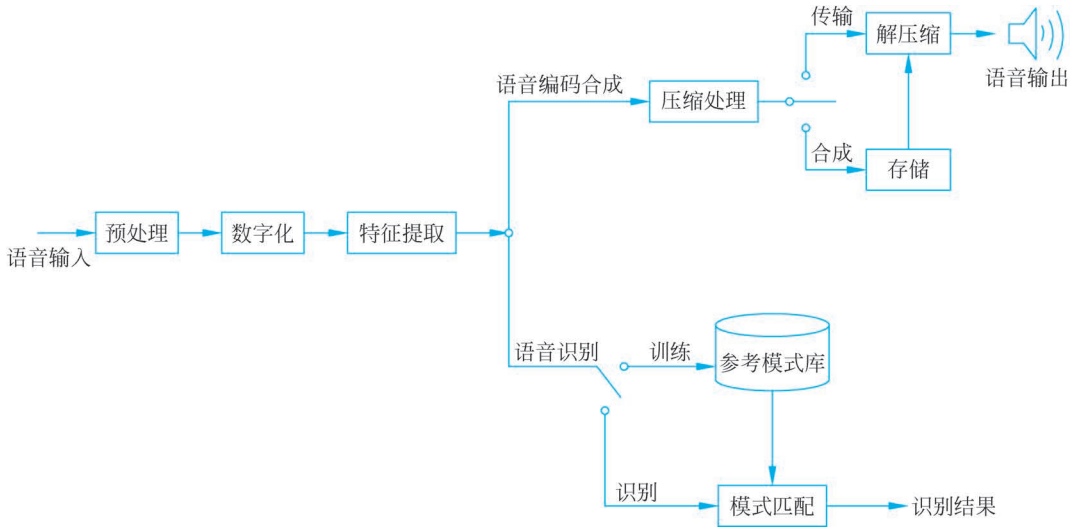


图 3.2　语音信号处理的总体结构图

无论是语音识别，语音编码还是语音合成，输入的语音信号首先要进行预处理，对信号进行适当放大和增益控制，并进行反混叠滤波来消除工频信号的干扰；然后进行数字化，将模拟信号转化为数字信号，便于计算机处理；接着进行特征提取，用反映语音信号特点的若干参数来代表语音。在此之后，根据任务的不同采取不同的处理办法。

语音信号处理技术已经渗透到生活的方方面面，从智能助手的便捷交互到安全领域的高级加密，它在不同行业展现出无限的可能性。深入探索这些应用，不仅可以帮助我们理解语音技术的实际效用，还能激发更多的创新思路。

3.1.2　语音识别技术

1. 语音识别的基本概念

语音识别技术（Automatic Speech Recognition，ASR）是一种利用机器对语音信号进行识别和理解，并将其转换成相应文本和命令的技术，其本质是一种模式识别，通过对未知语音和已知语音的比较，匹配出最优的识别结果。这项技术融合了生理学、声学、信号处理、计算机科学、模式识别、语言学和心理学等多个学科领域的知识，因此被视为一门交叉学科。

2. 语音识别技术的发展历程

20 世纪 50—70 年代（萌芽起步阶段）：1952 年，贝尔实验室研发了世界上第一个语音

识别系统 Audry，它可以识别 10 个英文数字发音。20 世纪 60 年代，计算机的应用推动了语音识别技术的发展，人们使用电子计算机进行语音识别，提出了一系列语音识别技术的新理论动态规划线性预测分析技术，较好地解决了语音信号产生的模型问题。代表性方法是动态时间规整(Dynamic Time Warping，DTW)，它依靠动态规划(Dynamic Programming，DP)技术解决了语音输入输出不定长的问题。20 世纪 70 年代，随着自然语言理解以及微电子技术的发展，语音识别研究取得了突破性进展。日本的 Sakoe(迫江)和 Chiba(千叶)的研究则展示了利用动态规划技术在待识语音模式与标准语音模式之间进行非线性时间匹配的方法；日本板仓的研究则提出了将线性预测分析技术加以扩展，将其用于语音信号特征抽取的方法。同时，这个时期还提出了矢量量化(Vector Quantification，VQ)，VQ 是将词库中的字、词等单元形成矢量量化的码本，作为模板，再用输入的语音特征矢量与模板进行匹配。总体而言，这一阶段主要实现了小词汇量、孤立词的语音识别。

20 世纪 80 年代—21 世纪初(发展阶段)：这一阶段的语音识别主要是以隐马尔可夫模型(Hidden Markov Model，HMM)为基础的概率统计模型为主，识别的准确率和稳定性都得到极大提升。经典成果包括 1990 年李开复等研发的 SPHINX 系统，该系统以 GMM-HMM(Gaussian Mixture Model-Hidden Markov Model)为核心框架，是有史以来第一个高性能的非特定人、大词汇量、连续语音识别系统。GMM-HMM 结构在相当长时间内占据语音识别系统的主流地位，并且至今仍然是学习、理解语音识别技术的基石。

21 世纪至今(应用落地阶段)：这一阶段的语音识别建立在深度学习基础上，得益于神经网络对非线性模型和大数据的处理能力，取得了大量成果。2009 年，Mohamed 等提出深度置信网络(Deep Belief Network，DBN)与 HMM 相结合的声学模型在小词汇量连续语音识别中取得成功。2012 年，深度神经网络与 HMM 相结合的声学模型 DNN-HMM 在大词汇量连续语音识别(Large Vocabulary Continuous Speech Recognition，LVCSR)中取得成功，掀起利用深度学习进行语音识别的浪潮。此后，以卷积神经网络(Convolutional Neural Network，CNN)、循环神经网络(Recurrent Neural Network，RNN) 等常见网络为基础的混合识别系统和端到端识别系统都获得了不错的识别结果和系统稳定性。迄今为止，以神经网络为基础的语音识别系统仍旧是国内外学者的研究热点。

3. 语音识别的工作原理

从语音识别系统的构成来讲，一套完整的语音识别系统包括预处理、特征提取、声学模型、语言模型以及搜索算法等模块，具体的语音识别流程如图 3.3 所示。

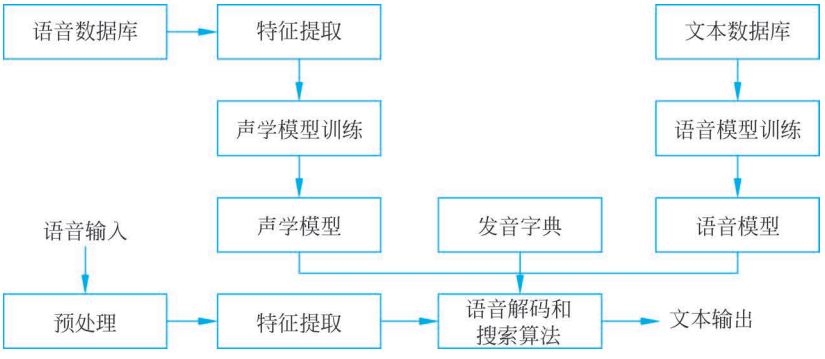


图 3.3　语音识别系统结构图

1）预处理

预处理包括预滤波、采样、模/数转换、预加重、分帧加窗、端点检测等操作。其中，信号分帧是将信号数字化后的语音信号分成短时信号，作为识别的基本单位。这主要是因为语音信号是非平稳信号，且具有时变特性，不易分析，但其通常在短时间范围（一般为 10～30ms）内特性基本不变，具有短时平稳性，可以用来分析其特征参数。

2）特征提取

通常在进行语音识别之前，需要根据语音信号波形提取有效的声学特征。特征提取的性能对后续语音识别系统的准确性极其关键，因此需要具有一定的鲁棒性和区分性。目前语音识别系统常用的声学特征有梅尔频率倒谱系数（Mel-Frequency Cepstrum Coefficient，MFCC）；感知线性预测（Perceptual Linear Predictive，PLP）系数、线性预测倒谱系数（Linear Prediction Cepstral Coefficient，LPCC）、梅尔滤波器组系数（Mel Filter Bank，Fbank）等。

3）声学模型

声学模型是整个语音识别系统中最重要的部分，只有学好了发音，才能顺利和发音词典、语言模型相结合，得到较好的识别性能。GMM-HMM 是最为常见的一种声学模型，该模型利用 HMM 对时间序列的建模能力，描述语音如何从一个短时平稳段过渡到下一个短时平稳段。深度学习的兴起为声学建模提供了新途径。

4）语言模型

语言模型是用来预测字符（词）序列产生的概率，判断一个语言序列是否为正常语句。随着深度学习的发展，语言模型的研究也开始引入深度神经网络。

3.1.3 声纹识别技术

声纹识别（Voiceprint Recognition），是一种通过分析语音信号中的个性化特征来识别或验证说话人身份的生物识别技术。每个人的发声器官（声带、口腔、鼻腔等）和发音习惯具有独特性，导致语音信号中隐含的声学特征（如基频、共振峰、频谱包络等）具有个体差异性。这些特征被称为“声纹”，类似于指纹的独特性。

1. 声纹特征

（1）交互性：声音是唯一可双向传递信号的生物特征，既可以接收信息，也可以发出信息，实现交互。

（2）便捷性：声音是唯一周边无死角的生物特征，可以实现非接触式采集，方便使用。

（3）变化性：声音是高可变性与唯一性的完美统一。没有两个声音是完全一样的，但里面所蕴含的信息，比如你是谁、你的年龄、你的情感等信息却都是唯一确定的。这种高可变性和唯一性的完美统一使得语音信号自身就具备了很强的防攻击能力。

（4）丰富性：声音有“形简意丰”的特点，它虽然只是一个一维信号，但是蕴含着丰富的信息。在一段语音中，除了包含说话人信息外，还包含内容、语种、性别、情绪、年龄，甚至包含出生地、身体健康状况等丰富的信息。

2. 声纹识别技术的原理

声纹识别和语音识别在原理上一样，都是通过对采集到的语音信号进行分析和处理提取相应的特征或建立相应的模型，然后据此做出判断。但二者的根本目的，提取的特征、建

立的模型是不一样的。声纹识别试图寻找的是区别每个人的个性特征，而语音识别则是侧重于对话者所表述的内容进行识别。简而言之，语音识别关心说的是什么，声纹识别关心是谁说的，声纹识别通常又称作说话人识别。

3. 声纹识别技术的分类

根据实际场景需求的区别，《2019 中国声纹识别产业发展白皮书》中将声纹识别技术细分为 4 类，如表 3.1 所示。

表 3.1　声纹识别的技术分类

技术分类	技术特点
声纹确认	即给定一个说话人的声纹模型和一段只含一名说话人的语音，判断该段语音是否是该说话人所说
声纹辨认	即给定一组候选说话人的声纹模型和一段语音，判断该段语音是哪个说话人所说
声纹检出	即给定一个说话人的声纹模型和一些语音，判断目标说话人是否在给定的语音中出现
声纹追踪	即给定一个说话人的声纹模型和一些语音，判断目标说话人是否在给定的语音中出现，若出现，则标示出对话语音中目标说话人所说的语音段的位置

3.1.4　语音合成技术

1. 语音合成的基本概念

语音合成，又称文语转换（Text-to-Speech，TTS）技术，它涉及声学、语言学、数字信号处理、计算机科学等多个学科技术，是中文信息处理领域的一项前沿技术，解决的主要问题是将文字信息转化为可听的声音信息，是一种让机器模仿人类说话者发出类似人的语音的技术。这项技术在现代科技社会中广泛应用，如广播电视、网络视听等。在传统的语音合成技术中，需要先录制一段人工语音，然后通过计算机算法将其转换为人工合成语音。随着人工智能技术的发展，语音合成技术也得到了快速发展，其应用场景也越来越广泛。

2. 语音合成技术的发展历程

语音合成技术的研究距今已有两百年历史，在第二次工业革命之前，语音的合成主要以机械式的音素合成为主。1779 年，德裔丹麦科学家克里斯蒂安·戈特利布·克拉岑斯坦建造了人类的声道模型，使其可以产生 5 个长元音。1791 年，沃尔夫冈·冯·肯佩兰添加了唇和舌的模型，使其能够发出辅音和元音。贝尔实验室于 20 世纪 30 年代发明了声码器（Vocoder），将语音自动分解为音调和共振，此项技术由荷马·达德利改进为键盘式合成器，并于 1939 年纽约世界博览会展出。语音合成技术的发展大致可以分为以下几个阶段。

（1）拼接合成阶段。

最早的语音合成系统采用录音单元拼接的方式，通过拼接预先录制的音素或音节来生成语音。这种技术最大限度地保留了原始发音人的音质，自然度和清晰度都很高，达到人们能够接受的水平。但这样直接拼接的方法导致语音听起来人工、生硬，韵律修饰导致边界处明显不连续。拼接处容易产生意想不到的错误，合成效果不稳定，音库容量大，构建周期长，可扩展性太差，不适宜作为嵌入式应用。

（2）参数合成阶段。

通过建立声学模型来描述语音的频谱特征，如共振峰频率等参数，再用这些参数驱动声码器合成语音。代表性方法有共振峰合成（formant 合成）和基于隐马尔可夫模型的参数合

成（HMM-based）。

（3）统计参数合成阶段。

随着计算机处理速度和存储容量的不断提升，语音合成技术也快速发展。20 世纪 90 年代，人们提出了基于统计参数的语音合成方法，采用统计模型如隐马尔可夫模型来建模语音参数的分布，能够生成更自然的语音。这种方法提出了语音合成十分重要的三个模块：语言模型、声学模型和声码器，如图 3.4 所示。其中，语言模型的任务是通过自然语言处理的技术将输入文本提取为语言特征，这些特征具有后端声学模型所需要的语言学信息。声学模型负责将语言特征转换为声学特征，再由单独的声码器完成声学特征到原始语音波形的转换。

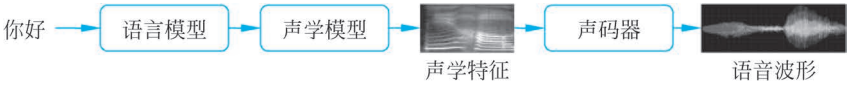


图 3.4　统计参数语音合成技术流程图

（4）深度学习阶段。

进入 21 世纪，随着深度学习技术的发展，语音合成技术快速发展。利用深度神经网络直接从文本特征映射到声学特征，大幅提升了合成语音的自然度和表现力，具体的技术流程如图 3.5 所示。代表性方法有 WaveNet、DeepVoice 等。2010 年，科大讯飞公司成功研发出首个基于深度学习的语音合成系统——讯飞语音合成系统。该技术使用了深度神经网络模型，能够实现更加自然流畅的语音合成效果。此后，科大讯飞公司在语音合成领域取得了重大突破，相继推出了“讯飞智能语音合成系统”和“讯飞混合语音合成系统”等多个系统。2017 年，百度公司发布了首个基于深度学习的语音合成系统 DeepVoice。该系统利用神经网络模型实现语音合成，具有较高的语音自然度和情感表达能力。

图 3.5　深度学习语音合成技术流程图

（5）端到端神经网络阶段。

采用端到端的神经网络架构，可以直接将语音信号映射到文本信息，无须手动提取语音特征或训练隐马尔科夫模型等传统方法，简化了系统的设计与实现，提高了识别性能。代表性方法有 FastSpeech、VITS 等。2020 年，阿里巴巴自然语言处理实验室提出了 Meta-VoiceGAN 模型，该模型采用基于 GAN 的方法，通过学习语音信号与语音特征之间的映射关系实现了高保真度的语音合成效果。2021 年，京东 AI 实验室发布了“京东流式语音合成技术”，该技术采用了基于 Transformer 的神经网络模型，结合预训练和微调等技术，能够实现更加自然流畅的语音合成效果，并具有较高的适应性和灵活性。目前，我国越来越多的科研单位大力投入 AI 语音合成的技术开发当中，未来技术发展和应用空间极为广阔。

3. 语音合成的工作流程

语音合成的工作流程主要包括文本分析和语音合成两大模块。文本分析阶段涉及将输入的文本转换为语音合成的内部表示，包括文本预处理、分词与词性标注、语法分析、韵律分析等。语音合成阶段则是将这些内部表示转换为声音波形，包括声学特征生成、波形合成、后处理等。这两个阶段紧密相连，高效地完成从文本到语音的转换任务，具体流程如图 3.6 所示。

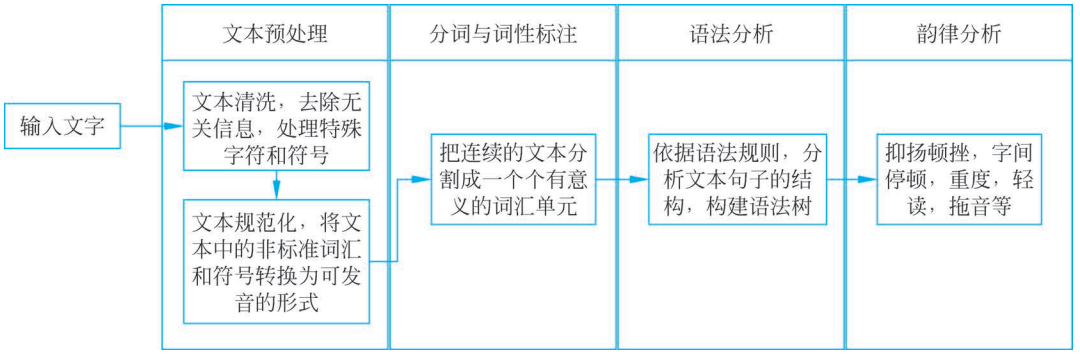


图 3.6　文本分析基本流程

1）文本分析

（1）文本预处理。

将原始文本转换为标准化的文本格式，包含无关信息、特殊符号等情况。此步骤首先对文本进行清理，去除无关的标点符号、特殊字符，统一文本格式，例如将全角字符转换为半角字符，去除文本中的多余空格、标点符号、HTML 标签，将数字（如“123”）转换为可发音的形式（如“一百二十三”）等。

（2）分词与词性标注。

把连续的文本分割成一个个有意义的词汇单元，如将“我喜欢学习人工智能课程”分词为“我”“喜欢”“学习”“人工智能”“课程”。确定每个分词后的词汇在句子中的词性，如名词、动词、形容词等。对于“我喜欢学习人工智能课程”这句话，“喜欢”为动词，“人工智能”和“课程”被标注为名词。词性标注有助于理解词汇在句子中的语法作用和语义关系，为后续语法和语义分析提供支持。

（3）语法分析。

依据语法规则分析文本句子的结构，构建语法树。例如对于“我喜欢人工智能课程”，语法分析可确定“我”是主语，“喜欢”是谓语，“人工智能课程”是宾语，其中“人工智能”是名词短语，“人工智能”修饰“课程”，表明课程的类型或内容。

（4）韵律分析。

人类在语言表达的时候总是附带语气与感情，语音合成的音频是为了模仿真实的人声，所以需要对文本进行韵律预测，如什么地方需要停顿，停顿多久，哪个字或者词语需要重读，哪个词需要轻读等，实现声音的高低曲折，抑扬顿挫。

2）语音合成

（1）声学特征生成。

根据文本分析结果确定语音的韵律特征，包括语调、语速、重音等。如陈述句语调通常较为平稳，而疑问句语调则会在句末上扬，重音也会根据语义重点进行标注。基于韵律规划信息，结合声学模型生成语音的声学参数，如基频、共振峰、时长等。这些参数决定了语音的音高、音色、音长等物理特性，确保合成的语音更加自然流畅。

（2）波形合成。

利用生成的声学参数，选择合适的波形合成方法，将其转换为实际的语音波形。常见的波形合成方法包括拼接合成、参数合成、深度学习合成等。

（3）后处理。

对合成的语音波形进行优化，包括降噪处理，去除可能存在的背景噪音或合成过程中产生的杂音；调整音量，使语音音量适中且保持一致；还可能进行音色调整，让语音听起来更加自然、舒适，最终输出可供播放的高质量语音。

4. 当前语音合成的主流方法

1）WaveNet

WaveNet 是 2016 年谷歌公司 DeepMind 开发的一种深度神经网络，旨在生成人类自然语音。与传统的语音合成方法相比，WaveNet 具有显著的优势，能够生成听起来更真实、更自然的语音，擅长语音合成、音乐生成以及音效合成等任务。WaveNet 的工作原理是通过使用真实语音记录训练的神经网络来直接模拟波形，从而生成类人声音。这是一种概率性和自回归性的生成方式，意味着对于每个预测的音频样本，其分布都基于前面的样本分布。这种技术使得 WaveNet 能够生成具有连续性和自然性的语音，而不仅仅是单个音素或音节。WaveNet 具有如下优点。

（1）高质量语音。

WaveNet 生成的语音听起来非常自然，几乎与人类录制的语音无法区分。这是因为 WaveNet 直接模拟波形，而不是简单地复制或合成已有的语音样本。

（2）连续性。

WaveNet 能够生成连续的语音样本，这意味着它能够模拟出流畅的语音流，而不会出现音素之间的断裂或不连续性。

（3）自然度。

由于 WaveNet 是基于真实语音数据训练的，因此生成的语音具有很高的自然度。这使得 WaveNet 在语音合成和语音识别领域都有广泛的应用。

2）Tacotron 系列

2017 年，谷歌公司提出了基于端到端的 Tacotron 语音合成模型，Tacotron 一经出现，便立刻成了端到端语音合成技术的标杆。Tacotron 训练使用的是带标注的数据，即文本和语音配对的数据，它不需单独使用一个模块来提取语言学特征。该模型包括声学模型和声码器两部分，从文本直接生成音频波形。这解决了很多前端文本分析的复杂问题，简化了语音合成模型的网络结构，同时还提高了语音合成的速度及质量。

因为 Tacotron 是一个完全的端到端模型，直接将文本特征映射到声学特征上，从被提出后就一直受到研究人员广泛的关注，研究人员在最初的 Tacotron 版本基础上提出了各种各样的改进版本。Tacotron 也得到了谷歌的关注，并在 Tacotron 被提出不久后对其进行了深入研究，把 Tacotron 的网络结构进行优化，并且把其和 WaveNet 进行组合，提出了 2.0 版本 Tacotron2 模型，提升了语音的自然度和质量。Tacotron 系列模型在中文语音合成中的应用表现出色，尤其在处理方言和非标准发音方面。

3）FastSpeech 系列

FastSpeech 系列是语音合成领域的重要成果，它由微软研究院与浙江大学研究团队于 2019 年首次发表在国际机器学习顶会 NeurIPS 上，标志着非自回归语音合成模型的重大突破。这一模型的诞生旨在解决传统神经网络端到端语音合成模型（如 Tacotron2）存在的诸多问题，如推理速度慢、合成语音不稳定（存在跳字或重复字现象）以及缺乏可控性（难以对

语速、韵律等进行有效控制)。FastSpeech 基于 Transformer 架构构建,采用非自回归方式生成语音,这一创新性的思路为语音合成技术开辟了新的方向。FastSpeech 模型具有以下优势:

(1) 推理速度极快。

能依据音素时长预测结果对源音素序列进行扩展,实现并行生成梅尔频谱图。与自回归 Transformer TTS 模型相比,其梅尔频谱图生成速度快了 270 倍,端到端语音合成速度快 38 倍,这使得实时语音合成得以轻松实现。

(2) 语音输出稳定且高质量。

它通过从基于编码器-解码器的教师模型提取注意力对齐信息,用于音素时长预测,长度调节器据此对音素序列合理扩展,避免了跳字和重复字问题,让生成的语音更加流畅自然,与人类真实语音相似度更高。

(3) 强大的可控性。

能灵活调节语速,通过调整长度调节器参数,加快语速就缩短音素持续时间,减慢语速则延长音素持续时间。还能通过调节句子中的空格字符持续时间控制单词间停顿,对合成语音韵律进行调整,以适应不同应用场景和表达需求。

FastSpeech 系列凭借“速度+质量+可控性”的三重优势重塑了语音合成领域的技术格局。从初代的速度革命到 FastSpeech2 的质量突破,该系列持续推动行业向更智能、更人性化的语音交互迈进,成为支撑 AI 语音生态的核心技术之一。FastSpeech 系列模型的应用非常广泛。在智能语音助手方面,像苹果 Siri、亚马逊 Alexa、小米小爱同学等,FastSpeech 的快速推理速度和高质量语音合成能力,使智能语音助手能迅速响应用户指令,并以自然流畅的语音回答,极大提升了交互体验。语音导航系统也离不开 FastSpeech,它能快速把导航信息转化为清晰准确的语音提示,根据路况和导航场景合理调整语速和重点,保障用户及时理解导航信息。

3.2　语音处理技术的典型应用

语音处理技术广泛应用于通信、语音识别、语音合成、音频处理等领域,提高通信系统的效率和用户体验,在安全验证和多模式通信方面发挥着关键作用,为科技的不断进步提供动力。随着人工智能的飞速发展,语音处理技术逐步演变为结合深度学习等人工智能技术的智能语音识别。智能语音交互是人们接触智能语音最普遍的渠道,从手机语音助手、家庭智能音箱、智能耳机、智能电视、故事机到智能车载等等。下面主要以智能语音助手、语音导航系统、语音交互教育产品为例介绍语音处理技术的典型应用。

3.2.1　智能语音助手

智能语音助手是用于终端的语音控制程序,通过智能对话与即时问答的智能交互让智能机器助手帮助完成用户指派的任务。2011 年,第一款手机语音助手 Siri 伴随 iPhone 4S 亮相,各大厂商纷纷入局。从 2017 年下半年开始,通过开放语音生态系统进行产业内合作,语音助手向家居、车载、可穿戴设备等领域不断延伸和迁移,构建出全产业生态链。

1. 智能语音助手的发展历程

智能语音助手从简单的语音识别发展到深度学习驱动的智能助手，如今正在向更高阶的类人交互 AI 迈进，未来将更加智能化、个性化和无缝地融入人类生活。智能语音助手的发展大致可以分为以下几个阶段。

20 世纪 60—80 年代（初期探索阶段）：1962 年，IBM 推出首个语音识别系统 Shoe box，可识别 16 个英文单词和数字，标志着语音技术首次应用于指令识别。1971 年，美国国防部研究所资助了为期五年的语音理解研究项目，推动了语音识别技术的一次重大发展。当时的 IBM、卡内基-梅隆大学、斯坦福大学等学术界和工业界的顶尖研究机构都加入了语音识别技术的研究。卡内基-梅隆大学开发的 Harpy 语音识别系统能够识别 1010 个单词，在发展初期取得了大词汇量孤立词识别方面的实质性进展，但主要用于文字转录，准确率较低。1984 年，IBM 发布的语音识别系统在 5000 个词汇量级上达到了 95%的识别率。

20 世纪 90 年代—21 世纪 10 年代（深度学习驱动的产品智能化阶段）：20 世纪 90 年代，语音助手采用隐马尔可夫模型（HMM）与高斯混合模型（GMM）成为主流，提升语音识别准确率。代表产品如 Dragon Systems 的 Dragon NaturallySpeaking（1997 年），其首次实现大词汇量连续语音识别，并进入消费市场。20 世纪初，深度学习（DNN）、卷积神经网络（CNN）、循环神经网络（RNN）、Transformer 模型飞速发展，支持多轮对话且具备一定的情境理解能力，大幅提升了语音识别和语义理解能力。2011 年，苹果推出 Siri，标志着语音助手进入移动互联网时代。2012 年，谷歌公司推出 Google Now，强调预测性搜索。2014 年，微软公司推出 Cortana，整合 Windows 生态。2016 年，Google Assistant 发布，整合深度学习，提供更精准的回答。2017 年，亚马逊的 Alexa 和 Echo 音箱普及，语音助手进入智能家居领域。2018 年，百度公司推出 DuerOS，小米推出“小爱”同学，国内语音助手快速发展。

21 世纪 20 年代至今（大模型与多模态 AI 阶段）：大语言模型（LLM）如 GPT-3、GPT-4 的兴起，使语音助手进入多轮交互、上下文理解、个性化推荐的时代。AI 可以生成更自然的对话，甚至能处理编程、创意写作等任务。语音助手开始支持多模态交互（语音＋文本＋图像），与 AR/VR、元宇宙等技术结合。2023 年，OpenAI 公司发布 GPT-4＋Whisper 语音系统，具备流畅语音对话和上下文记忆功能。2024 年，国内企业如百度（文心一言）、科大讯飞、阿里巴巴等相继将大模型集成至语音助手系统。

2. 智能语音助手的工作流程

（1）语音唤醒。

通过语音（指定词语）唤醒设备，这里的“唤醒”指的是让设备从待机状态进入工作状态，开始对用户的话语进行监听、识别与回应。其中唤醒词分为固定唤醒词、自定义唤醒词，固定唤醒词一般为语音助手的名称（如“小度小度”），而自定义唤醒词一般为用户定义的指令（如“打开空调”）。

（2）语音识别。

通过麦克风阵列收集用户语音信号，将声音的模拟信号转换为数字信号，进行降噪、滤波、回声消除等预处理操作，提升语音信号质量。将预处理后的语音信号转化为计算机可处理的特征向量，并与预先训练好的声学模型进行比对，计算出最可能的音素序列。最后结合语言模型对声学模型输出的音素序列进行解码，得到最终的文本结果。

(3) 语义理解。

对语音识别得到的文本进行清理、分词、词性标注等操作，将文本转化为适合计算机处理的格式。通过机器学习算法或深度学习模型，如支持向量机、卷积神经网络等对文本进行分析，判断用户的意图，如查询天气、设置闹钟、播放音乐等。确定意图后，从文本中提取关键信息，例如查询哪个城市的天气，将用户的问题与已有的知识体系进行关联和匹配，获取更准确和全面的信息。

(4) 指令执行。

根据语义理解的结果确定需要执行的任务，并将任务分配给相应的模块或功能组件。如果需要获取外部信息或执行特定操作，如查询天气、发送短信等，语音助手会调用相应的API接口或应用程序来完成任务。对于与智能设备连接的语音助手，还可以直接控制设备的运行状态，如调节灯光亮度、控制家电开关等。

(5) 语音反馈。

根据指令执行的结果生成要反馈给用户的文本内容。利用文本转语音技术(TTS)，将文本转换为自然流畅的语音信号。TTS技术通常基于深度学习模型，如WaveNet、Tacotron等，能够生成接近人类语音的声音。通过扬声器将合成的语音播放出来，反馈给用户。

3.2.2　语音导航系统

语音导航是以语音识别、语音编解码为代表的智能语音技术，该技术应用在车载领域，实现车内语音声控操作，可改变汽车现有的人机信息交流方式，极大提升了驾驶的便捷性和安全性。其核心功能包括语音指令输入、路线规划语音播报、实时路况语音反馈等，均依赖于语音处理技术的支持。

(1) 语音指令输入。

借助语音识别技术，将用户的语音转化为机器可理解的指令。当用户说出如“导航到中心公园”等指令时，语音导航设备首先对输入语音进行预处理，包括降噪、增益等操作，以提高语音质量。随后，利用声学模型将语音信号转换为对应的音素序列，再通过语言模型分析音素序列的语法和语义，确定用户的意图。

(2) 路线规划语音播报。

借助语音合成技术，在系统规划好从起点到目的地的路线后，将文字形式的路线信息(如“前方500米右转进入XX路”)通过语音合成转化为语音输出。语音合成技术从文本预处理开始，对文字进行分词、词性标注等分析，确定每个词汇的发音和韵律信息。接着，依据声学模型生成对应的声学参数合成语音波形。为了使播报更自然、易懂，还会融入情感合成技术，根据路况和导航信息调整语音的语调、语速和重音。例如，在复杂路口转弯提示时，加重语气，并适当放慢语速，引起驾驶者注意，让驾驶者在无须分心查看屏幕的情况下清晰获取路线指引，在提高出行效率的同时保障驾驶专注度与安全性。

(3) 实时路况语音反馈。

综合运用了语音识别、语音合成以及数据处理等技术。语音导航系统通过与交通数据中心实时连接，获取道路拥堵、事故、施工等路况信息。当路况发生变化，影响原定路线时，系统自动分析并重新规划路线。同时，利用语音合成技术，将新的路况信息和路线调整建议

以语音形式反馈给用户。例如，“前方路段拥堵，建议您选择 XX 备用路线，预计可节省 15 分钟行程时间”。在此过程中，语音识别技术可用于用户对路况信息的进一步询问，如“这条备用路线的具体情况”，系统识别指令后，经数据处理和分析，再通过语音合成给予准确回复。这种实时语音反馈让驾驶者及时了解路况，灵活调整路线，避免拥堵，提高出行效率，减少因交通堵塞带来的烦躁情绪，间接提升驾驶安全性。

随着人工智能技术不断迭代，语音导航的语音交互将愈发精准智能。一方面，语音识别技术会进一步优化，能在更复杂环境下准确识别语音指令。例如，在嘈杂的车内环境，结合多麦克风阵列降噪技术与先进的语音识别算法，即便车内播放音乐、乘客交谈，系统也能精准捕捉驾驶者的语音，理解复杂语义。另一方面，自然语言处理能力提升，让语音导航不仅能理解指令，还能像人类对话般交流，根据用户过往出行习惯、偏好主动提供个性化建议。

3.2.3 语音交互教育产品

在智能教育领域，AI 课堂的建设进入快车道：一是解决家校之间、线上线下之间学习资源互通的问题，二是通过多模态识别收集课堂学情信息，并做数据精准分析，因此通过语音转录、语音识别等技术实现授课语音转录为文字、利用多模态识别进行课堂质量监测不可或缺。另一方面，在线教育竞争呈白热化态势，用技术解决教育资源的复用、增加学习交互体验感等诉求也促进了智能语音技术在线上口语测评、虚拟教师等领域的应用。智能语音在教育领域的主要应用如下。

（1）语音转录丰富教学模式：通过语音识别实时转写教师讲课的语音为文字，可在授课视频中嵌入字幕，并进行关键词和知识点的快速定位，应用于直播课、小班课、互动课堂。

（2）语音算法助力课堂质量监测：利用静音检测、语速检测，结合计算机视觉等多模态算法，自动化监测上课互动情况和教学质量。

（3）虚拟教师互动教学：通过语音合成＋VR 技术可以打造虚拟的名师形象，通过亲切的语音、动作、文字等方式与学生互动。

（4）口语测评：可对语音的完整性、韵律节奏及语义、语法进行评测等综合打分，有些产品涉及发音纠正功能，中文测评还可覆盖轻音、儿化音等汉语语音特征，可用于日常口语学习及新中/高考口语机考。

3.3 语音处理实践

请使用 Python 的第三方库 SpeechRecognition 库编写代码，分别实现麦克风实时语音采集识别和识别本地音频文件。

1. 任务准备

（1）使用 pip 命令安装第三方库 SpeechRecognition、pyaudio、chardet、baidu-aip，安装过程以 SpeechRecognition 为例，如图 3.7 所示。

SpeechRecognition 是一个功能强大且易于使用的 Python 第三方库，用于执行语音识别任务。它为开发者提供了统一的接口，能够与多种语音识别引擎集成，方便实现从语音到文本的转换。该库具有以下特点。

① 支持多种语音识别引擎：支持 Google Web Speech API、Microsoft Bing Voice

Recognition、IBM Speech to Text 等多种流行的语音识别服务，开发者可以根据需求和场景选择合适的引擎。

② 跨平台兼容性：可以在 Windows、macOS、Linux 等多种操作系统上使用，只要 Python 环境能够正常运行，就能利用该库进行语音识别开发。

③ 多种音频源支持：能够从麦克风、音频文件（如 WAV、AIFF 等常见格式）获取音频数据，为不同的应用场景提供便利。

```
  Downloading idna-3.10-py3-none-any.whl (70 kB)
     |████████████████████████████████| 70 kB 180 kB/s
Collecting certifi>=2017.4.17
  Downloading certifi-2025.1.31-py3-none-any.whl (166 kB)
     |████████████████████████████████| 166 kB 218 kB/s
Collecting urllib3<3,>=1.21.1
  Downloading urllib3-2.0.7-py3-none-any.whl (124 kB)
     |████████████████████████████████| 124 kB 243 kB/s
Installing collected packages: charset-normalizer, idna, certifi, urllib3, requests, SpeechRecognition
Successfully installed SpeechRecognition-3.9.0 certifi-2025.1.31 charset-normalizer-3.4.1 idna-3.10 requests-2.31.0 urllib3-2.0.7
```

图 3.7　安装第三方库 SpeechRecognition

pyaudio 库用于处理音频输入和输出，为开发者提供了与底层音频设备交互的便捷方式。chardet 是一个用于自动检测文本编码的 Python 库。它通过分析文本的字节模式推测出最可能的字符编码（如 UTF-8、GBK、ISO-8859-2 等），帮助开发者处理未知编码的文本数据。baidu-aip 库是百度公司 AI 开放平台的官方 PythonSDK，用于快速调用百度 AI 服务（如 OCR、语音识别、自然语言处理等）。

```
PS C:\Users\Terry\Desktop > pip install SpeechRecognition
```

温馨提示：除了能使用 pip 命令安装第三方库，还可以直接通过 PyCharm 软件中的“设置”，在相应的 Python 解释器中搜索所需的第三方库和相应版本安装，如图 3.8 所示。

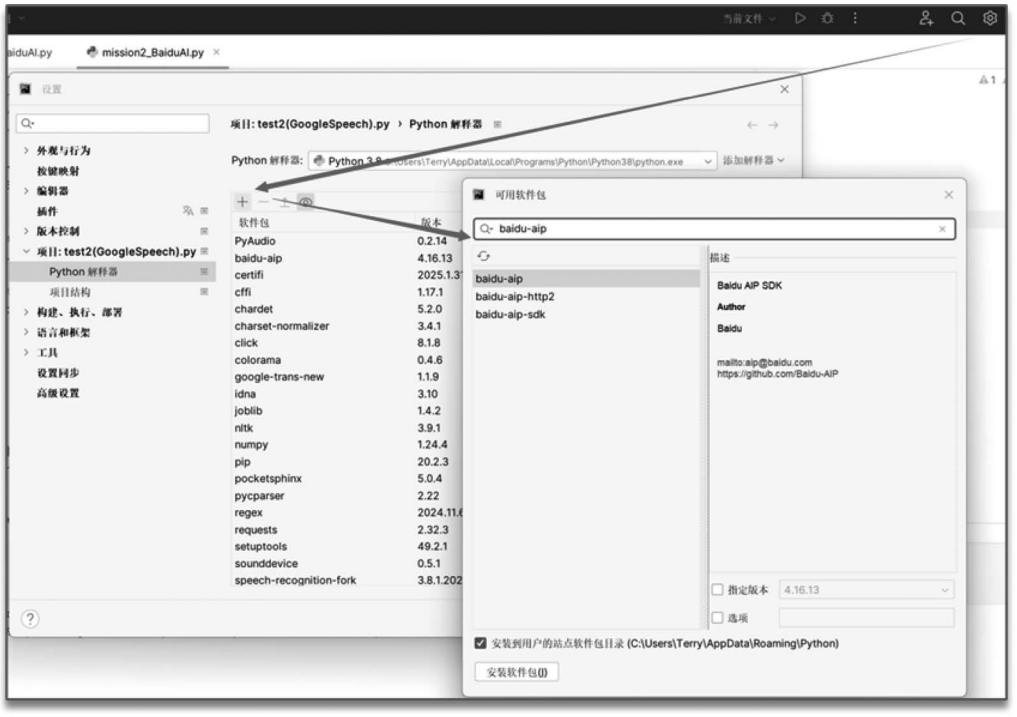


图 3.8　在 PyCharm 中直接安装第三方库

（2）获取百度 AI 的语音识别服务。

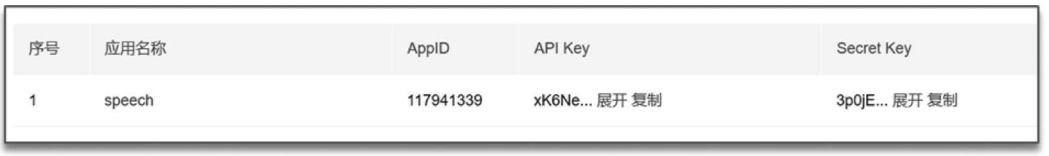
① 注册登录：访问百度公司 AI 开放平台官网地址(https://ai.baidu.com/)，使用本人手机号注册百度账号，如先前已用本人手机号注册过相关百度应用，可直接使用百度账号登录，并完成个人实名认证，便于后续使用百度提供的 AI 服务。

② 创建应用——调用百度 AI 服务：通过控制台左侧导航选择产品导览→人工智能，自行选择“语音技术”人工智能服务项。进入语音技术控制台，单击使用指引模块的快速接入服务按钮。应用名称和应用描述应当尽量反映应用的实际用途，方便后续管理应用。在服务接口列表确保已勾选需调用的接口“短语音识别”和“实时语音识别”，勾选完毕后单击“立即创建”按钮。具体操作流程如图 3.9 和图 3.10 所示。

图 3.9　创建百度 AI 应用(1)

图 3.10　创建百度 AI 应用(2)

③ 获取 API Key 和 Secret Key：创建成功后，可以在应用列表页查看应用的 API Key 和 Secret Key，如图 3.11 所示。这是后续调用该应用内接口的凭证。



序号	应用名称	AppID	API Key	Secret Key
1	speech	117941339	xK6Ne... 展开 复制	3p0jE... 展开 复制

图 3.11　获取 API Key 和 Secret Key

2. 任务实施

(1) 编写以下代码，实现麦克风实时语音采集识别。

```
import speech_recognition as sr
# pip install chardet
# pip install baidu-aip
# 导入百度语音识别 SDK 客户端
from aip import AipSpeech
# 百度 AI 语音识别配置
APP_ID = '117778282'                                          # 替换你的 AppId
BAIDU_APP_KEY = 'LLUPeJLJoq7A52QIZNokYzvV'                    # 替换你的 App Key
BAIDU_APP_SECRET = 't5w61aaWek5rhWXLIaqcGmDscvEW0Cyy'         # 替换你的 Secret Key
client = AipSpeech(APP_ID, BAIDU_APP_KEY, BAIDU_APP_SECRET)
# 分析音频数据
def get_text(wav_bytes):
    result = client.asr(wav_bytes, 'wav', 16000,
                        # 语言类别
                        {'dev_pid': 1537, })
    try:
        text = result['result'][0]
    except Exception as e:
        print(e)
        text = ""
    return text
def recognize_speech():
    # 创建识别器实例
    recognizer = sr.Recognizer()
    # 配置麦克风
    with sr.Microphone() as source:
        print("正在调整环境噪声，请保持安静...")
        recognizer.adjust_for_ambient_noise(source, duration = 1)
        print("准备就绪，请开始说话...")
        while True:
            try:
                # 实时监听麦克风输入
                audio = recognizer.listen(source, timeout = 5, phrase_time_limit = 10)
                print("识别中...")
                audio_data = audio.get_wav_data(convert_rate = 16000)
                print('正在分析...')
                text = get_text(audio_data)
                # 调用百度语音识别 API
                # text = recognizer.recognize_google(audio,language = 'zh-CN')
                if text:
                    print(f"识别结果：{text}")
                else:
```

```
                print("没有识别到有效语音")
            except sr.WaitTimeoutError:
                print("检测超时,请重新说话...")
            except sr.UnknownValueError:
                print("无法识别语音")
            except sr.RequestError as e:
                print(f"请求错误：{e}")
            except KeyboardInterrupt:
                print("\n语音识别已停止")
                break
if __name__ == "__main__":
    # 验证配置信息
    if BAIDU_APP_KEY != 'LLUPeJLJoq7A52QIZNokYzvV' or BAIDU_APP_SECRET != 
't5w61aaWek5rhWXLIaqcGmDscvEW0Cyy':
        print("请先配置正确的百度 AI App Key 和 Secret Key")
    else:
        recognize_speech()
```

程序运行结果如图 3.12 所示。

```
C:\Users\Terry\AppData\Local\Programs\Python\Python38\python.exe D:\pythonProject2\mission1_BaiduAI.py
正在调整环境噪声，请保持安静...
准备就绪，请开始说话...
识别中...
正在分析...
识别结果：今天天气真不错。
```

图 3.12　实时语音识别程序运行结果

（2）编写以下代码，实现识别本地音频文件（.wav 格式）。

```
import base64
import urllib
import requests
import json
API_KEY = "LLUPeJLJoq7A52QIZNokYzvV"                    # 替换你的 App Key
SECRET_KEY = "t5w61aaWek5rhWXLIaqcGmDscvEW0Cyy"         # 替换你的 Secret Key
def main(file_path):
    url = "https://vop.baidu.com/server_api"
    # speech 可以通过 get_file_content_as_base64("C:\fakepath\test.wav",False) 方法获取
    payload = json.dumps({
        "format": "pcm",
        "rate": 16000,
        "channel": 1,
        "cuid": "V4tWTfRkV8q4dXH2DoZuxeFnto5KF5Qu",
        "speech": get_file_content_as_base64(file_path),
        "len": 92526,
        "token": get_access_token()
    }, ensure_ascii = False)
    headers = {
        'Content-Type': 'application/json',
        'Accept': 'application/json'
    }
    response = requests.request("POST", url, headers = headers, data = payload.encode("utf-8"))
    print(response.text)
def get_file_content_as_base64(path, urlencoded = False):
    """
    获取文件 Base64 编码
```

```
    :param path: 文件路径
    :param urlencoded: 是否对结果进行 urlencoded
    :return: Base64 编码信息
    """
    with open(path, "rb") as f:
        content = base64.b64encode(f.read()).decode("utf8")
        if urlencoded:
            content = urllib.parse.quote_plus(content)
    return content
def get_access_token():
    """
    使用 AK,SK 生成鉴权签名(Access Token)
    :return: access_token,或是 None(如果错误)
    """
    url = "https://aip.baidubce.com/oauth/2.0/token"
    params = {"grant_type": "client_credentials", "client_id": API_KEY, "client_secret":
SECRET_KEY}
    return str(requests.post(url, params=params).json().get("access_token"))
if __name__ == '__main__':
    # 上传的文件路径
    file = "D:\\pythonProject2\\test.wav"
    main(file)
```

程序运行结果如图 3.13 所示。

```
C:\Users\Terry\AppData\Local\Programs\Python\Python38\python.exe D:\pythonProject2\mission2_BaiduAI.py
{"corpus_no":"7479675596048391930","err_msg":"success.","err_no":0,"result":["快点过来。"],"sn":"858974328891741497683"}


进程已结束,退出代码0
```

图 3.13　识别本地音频文件程序运行结果

3.4　本章小结

本章系统性介绍了语音处理技术的基础理论与应用实践，构建了从基础概念到前沿技术的完整知识体系。在理论基础部分，首先阐述了语音信号的声学特征及其处理流程，随后深入解析了语音识别与语音合成两大核心技术，包括其技术演进历程、典型算法框架（如WaveNet、Tacotron、FastSpeech 等系列模型）以及完整的工作流程。在应用实践层面，本章选取了智能语音助手、车载语音系统和教育语音产品三大典型场景，详细分析了其系统架构、实现原理以及对社会生活的积极影响。为强化理论与实践的结合，最后通过 Python 的 SpeechRecognition 库实现了语音识别的基础应用开发，包括实时语音识别和本地音频文件处理，为读者的深入学习提供了实践切入点。