分类号_____ 密级_____

U D C _____ 编号 10736

西北师范大学

硕士学位论文

统计参数情感语音合成的研究

研 究 生 姓 名：郝东亮

指导教师姓名、职称：杨鸿武 教授

专 业 名 称：电子科学与技术

研 究 方 向：信号检测与处理

二〇一六年五月

硕士学位论文
M. D. Thesis

统计参数情感语音合成的研究

Research on statistical parametric emotional speech synthesis

郝东亮

Hao Dongliang

二〇一六年五月

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包括其他人已经发表或撰写过的研究成果，也不包含为获得西北师范大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：郝东亮　　日期：2016.5.30

关于论文使用授权的说明

本人完全了解西北师范大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

(保密的论文在解密后应遵守此规定)

签名：郝东亮　导师签名：杨鸿武　日期：2016.5.30

摘　要

随着语音合成技术的研究与发展，合成语音音质得到较大提升，但当前语音合成技术的研究仍以中性化语音为主，对情感语音合成的研究较少。人类生活对智能语音的需求不仅要涵盖基本的文字内容，还要承载丰富的情感信息，情感语音合成的研究将是智能语音研究领域的必然趋势。本文建立了一个多说话人的多种情感的情感语音语料库，针对汉语统计参数语音合成中的上下文相关标注生成，设计了一套包含 6 层上下文信息的标注格式，在此基础上，采用多说话人的情感语音数据和统计参数语音合成方法，利用说话人自适应训练算法训练了情感语音的声学模型，实现了情感语音的合成。论文的主要工作和创新如下：

1.建立了一个多说话人的多种情感的语料库。在专业录音棚中，采用诱发方式激发录音人的情感，并进行录音。录制了 7 个男性说话人和 7 个女性说话人的 11 种典型情感的情感语音数据，并以 Microsoft WAV 格式（单通道、16bit、16kHz 采样频率）进行保存。

2.实现了一种面向普通话统计参数语音合成的标注生成算法。针对汉语统计参数语音合成中上下文相关标注的生成，设计了一套包含 6 层上下文相关信息的标注格式。以声韵母做为语音合成的合成基元，利用基于隐 Markov 模型(Hidden Markov Model, HMM)的统计参数语音合成方法，通过对合成语音音质的主、客观评测，验证了不同上下文信息对合成语音音质的影响。实验结果表明，本文设计的上下文相关的 6 层标注格式能够满足情感语音合成的需求。

3.提出了一种利用多个说话人的多种情感训练语料，利用统计参数语音合成方法实现情感语音合成的方法。首先利用多个说话人的情感语音语料，通过说话人自适应训练(Speaker Adaptation Training, SAT)得到多个说话人情感语音的平均音模型，然后利用目标说话人的目标情感的训练语料，经过说话人自适应变换，得到目标说话人目标情感的声学模型，进而合成出目标说话人的目标情感语音。实验结果表明，本方法合成得到的情感语音具有较高的自然度和情感相似度。

关键词：情感语音合成；情感语料库；上下文相关信息；标注格式；说话人自适应训练；统计参数语音合成

Abstract

The quality of synthesized speech makes a remarkable improvement with the progress of speech synthesis technology. However, current researches of speech synthesis technology mainly focused on neutral speech synthesis. There is the lack of studies on emotional speech synthesis. The needs for intelligent voice in the human life not only cover basic textual information, but also carry a large number of emotional information. Therefore, the study on emotional speech synthesis will be the inevitable trend in the intelligent voice research. The thesis establishes an emotional speech corpus including a variety of emotions recorded by multi-speaker. Then a six-level context-dependent label format is designed for generating context-dependent labels of Mandarin statistical parametric speech synthesis. Speaker adaptive training algorithm is employed to train the emotional acoustic model with multi-speaker's emotional speech corpus to achieve statistical parametric speech synthesis. The main works and originalities of the thesis are as follows:

Firstly, the thesis establishes a multi-speaker speech corpus with 11 kinds of emotions. The thesis induces emotional state of speakers to record emotional speech corpus in a professional studio. The emotional speech corpus includes 11 kinds of typical emotions narrated by 7 male speakers and 7 female speakers where speech signal is saved in the Microsoft WAV format (single-channel, 16bit, 16k Hz sampling frequency).

Secondly, the thesis realizes a label generation algorithm for Mandarin statistical parametric speech synthesis. A six-level context-dependent label format is designed aiming at the generation of context-dependent labels for statistical parametric speech synthesis, which uses the initial and the final of Mandarin as the synthesis unit. A Hidden Markov Model (HMM) based statistical parametric speech synthesis system is adopted to synthesize the Mandarin speech. We evaluate the influences of the different label information on quality of synthesized speech by subjective evaluation and objective evaluation. Tests show that the designed six-level context-dependent label format meet the need of Mandarin emotional speech synthesis.

Finally, the thesis proposes a statistical parametric emotional speech synthesis method by using a HMM-based statistical parametric speech synthesis with multi-speaker's multi emotional training speech corpus. A set of average emotional acoustic model is trained by applying multi-speaker's emotional speech corpus to the speaker adaptive training (SAT) algorithm. The target speaker's emotional speech corpus is then used to perform the speaker adaptation transformation to obtain a target speaker's acoustic model with target emotion for synthesizing target speaker's emotional speech. Tests show that the synthesized emotional speech has good naturalness and emotional similarity.

Keywords: emotional speech synthesis; emotional corpus; context-dependent information; tagging format; speaker adaptive training; statistical parametric speech synthesis

目　录

第 1 章 前言

1.1 情感语音合成概述

随着人工智能技术的高速发展，智能语音系统、人机交互技术得到了广泛应用。语言作为人与人之间最重要的交流方式，语音信号处理技术逐渐成为当下研究人员的研究热点。人们迫切希望可以和计算机实现更简捷、更人性化的交流，这不仅仅局限于传统的手工输入操作和计算机显示功能，而是希望可以采用类似人与人沟通的方式实现人机智能交互，完成指令下发和思想传达。

在人类的语音信号中，一般包含两个层次的内容，一方面是文本中的字符信息，也就是语言中所传达出的文字内容，我们可以从中得到表述者所传达出的事件内容；另一方面是语音信号中所包含的副语言信息或超语言信息，这就相当语音信号所传达出的另外一部分内容，这些信息可以准确的传达出说话人的真正意图，在人与人沟通的过程中，往往会因为副语言信息或超语言信息的区别而传达出不同的意思，听者也会领悟到不同的信息。

近些年来，随着语音合成技术的不断发展，从最初的物理机理的语音合成方法、源-滤波器的语音合成方法，到目前日趋成熟的波形拼接的语音合成方法、统计参数的语音合成方法，以及当下研究正盛的基于深度学习的语音合成方法[1,2]，合成语音的音质得到明显改善。然而，传统的语音合成方法，研究人员仅仅实现了把书面文字、字符转换为简单的口语输出，却忽略了说话人在言语表达过程中所携带的情感信息，单纯的文本-语音转换不仅让听者感到语音交流的单调、乏味，同时合成语音也不足以反映出说话人所传达的情感信息，进而导致听者对言语内容理解的偏差。

因此，如何提高合成语音表现力，将成为情感语音合成技术研究的重要内容，也将是未来语音信号处理领域研究的必然趋势。国外，上个世纪 90 年代，美国 MIT 实验室通过计算机程序实现了仿声学和语言学的发音功能；1998 年，欧洲 PHYSTA 项目对情感语料库的构建进行开发，并对人类情绪在人机交互的应用技术展开研究；同年，日本 ATR 实验室 JST、CREST ESP 项目为实现最接近人类自然发音的语音合成也开始了基础研究；到 20 世纪 90 年代中后期，国外一些企业对情感语音合成系统的开发和实现逐渐展开，如 IBM、Dragon System、Nuance、Microsoft 等。国内，一些著名大学及科研院所逐渐开展对情感计算的理论和方法的研究，如清华大学、中科院等。目前，随着深度学习研究方法的逐渐成熟，IBM、Google、百度、中科院等知名企业和科研院所，也开始采用深度学习的研究方法对情感语音合成展开研究。

语音合成(Speech Synthesis)是通过计算机自动的把各种形式的文本信息转化为自然语音的过程，又称作文语转换(Text-To-Speech, TTS)[3]。近年来，基于隐Markov 模型的统计参数语音合成方法[4]成为主流，这种方法可通过说话人自适应变换[5,6]得到不同说话人多种情感的合成语音。目前实验室分别采用韵律修改的情感语音合成方法[7]和基于单情感说话人语音数据自适应训练的统计参数语音合成方法[8]，实现了情感语音合成，但合成效果仍需改善，为了进一步提高合成得到的情感语音的自然度和情感相似度，本文采用基于多个情感说话人语音数据的说话人自适应训练的统计参数语音合成方法实现情感语音合成。

当前，情感语音合成的研究成果不仅可以应用在一些传统语音合成应用领域，比如有声读物、软硬件系统提示音、声讯服务等，还可以广泛应用到信息智能查询、人机对话、语音 E-mail、游戏娱乐等领域。

1.2 情感语音合成的方法

1.2.1 基于波形拼接的情感语音合成

基于波形拼接的情感语音合成方法，是目前情感语音合成领域中比较常用的方法之一。如图 1.1 所示，采用这种方法，需要录制并构建一个包含不同情感的大型情感语音数据库，并尽量保证每种情感的语音数据相互独立；之后对输入的文本进行文本分析和韵律分析，根据分析的结果获得合成语音基本的单元信息，比如：音节、半音节、音素等；最后，根据得到的单元信息，在先前标注好的语料库中选取出最合适的语音基元，根据需求进行一定的修改和调整，最终经过波形拼接的方式得到目标情感的合成语音。其中合成基元的选取依据主要包括该基元所处的语境信息、韵律结构信息、声学参数信息。其优点是因为合成情感语音基元直接来源于原始语音库，所以合成的语音能够保持情感语音的特征，并具有较好的情感相似度。

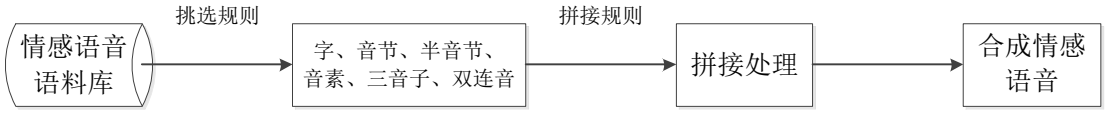


图 1.1 波形拼接合成流程图

采用波形拼接的情感语音合成方法，由于合成语音的基元均来自情感语料库的自然发音，合成语音包含真实的情感色彩，具有良好的情感相似度，理论上可以拼接得到任意情感色彩的合成语句。另外，由于受到情感语料库内容的制约，这种方法也存在着明显的缺陷。首先，这种方法只能得到情感语料库中所包含的相应的情感说话人有限的情感语音，对于语料库之外的其他说话人、其他文本内容起不到任何作用，扩展性较差；另一方面，这种方法得到的情感语音的自然度欠佳，音质也

不够清晰。如果想得到更多说话人、可懂度更高的情感语音，这就需要构建更大的语料库。除此之外，对拼接算法的优化、存储配置的调整等方面的要求也会提高。

早期波形拼接技术刚被提出时，音库的容量都比较小，且单元调整算法还不完善，因此合成语音的音质和连续性都不是很好，当对语音合成单元做出较大调整时，会对合成的语音音质产生明显影响。因此，当时这种方法的优势并没有完全体现出来，因为拼接合成的样本数比较少，也就是说用来合成的原始音库比较小，这种系统被称为样本语料库的波形修改拼接语音合成系统。后来，计算机的性能不断提升，语音音库的容量也随之扩充增大，因此可以更加准确地选择出所需要的合成基元，从而基本不需要对基元进行修改，可以让合成语音的自然度得到明显改善。由于所需语音基元都是从录制的语音语料库中直接选取的，合成语音能保持原说话人的语音特性。这种方法也被称为基于大语料库的基元选取波形拼接语音合成方法。

虽然基于大语料库的语音合成系统得到的语音自然度较高，合成语音清晰，也能够很好的体现目标说话人的发音特征，但是这种系统的性能并不稳定。有时候在语音基元的拼接的过程中，如果语句中的拼接点处存在过多的不连续的情况，便不能保证各个基元之间声学参数的连续性，并最终降低合成语音的音质。因此，上世纪 80 年代后期，E.Moulines 和 F.Charpentier 提出了基于波形拼接的基音同步叠加算法(Pitch Synchronous Overlap Add, PSOLA)[9,10]。这种方法先对待合成语音的韵律特征参数进行修改，并达到最优效果，再通过波形拼接方法，得到合成语音。基音同步叠加算法的提出，较大的改善了合成语音的自然度，有效的促进了波形拼接语音合成技术的发展。这种方法在小词汇的语音合成领域，比如旅游信息、天气预报等信息的合成得到极好的应用，当然，如果需要合成出更高质量的任意文本的语音，波形拼接方法还需要进一步改进和提高。

对于基元选取波形拼接情感语音合成来说，主要任务就是构建合适的情感语料库。Akemi[11]构建了包含高兴、悲伤、生气三种典型情感语料库，并利用该语料库采用波形单元选择方法，搭建了情感语音模型，并通过实验证明采用这种模型合成得到的情感语句可达到较高的识别率。语料库的构建过程包括语料设计、语音录制和音库制作这三个步骤。其中，在音库的制作过程中，需要对每个发音基元的韵律信息和边界信息进行标注。语音音库越大，语料库越充足，则合成的语音质量越好，但是要制作一个效果比较好的情感语音音库，工作量太大，周期也较长。早期的音库都是通过人工手动标注的，现在虽然可以通过程序实现语料的自动标注，但是其效果并不是很好，不是很稳定，因此，现在在语料库的构建上，大都选取单个或者比较少的说话人的语料，例如只选取一个男性或女性说话人的情感语料来构建音库，来减少音库标注和构建的工作量，但是这样合成的语音特性比较单一。除此之外，

在进行单元挑选时，现阶段的许多大语料库合成系统，当说话人、发音语种、发音情感发生变化时，需要对先前的单元挑选算法进行调整和优化，鲁棒性不高。

1.2.2 基于韵律特征的情感语音合成

韵律特征是情感信息传达必不可少的一部分，人在说话的过程中，言语中所携带的语气、情调都和韵律特征相关[12]。采用波形拼接的情感语音合成方法，可以通过扩大情感语音语料库，来提升合成语音的情感特征，但这种方法对韵律特征的控制能力是很有限的。韵律合成流程如图 1.2 所示。

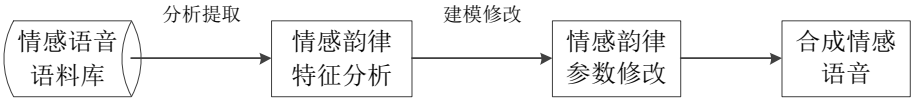


图 1.2 韵律合成流程图

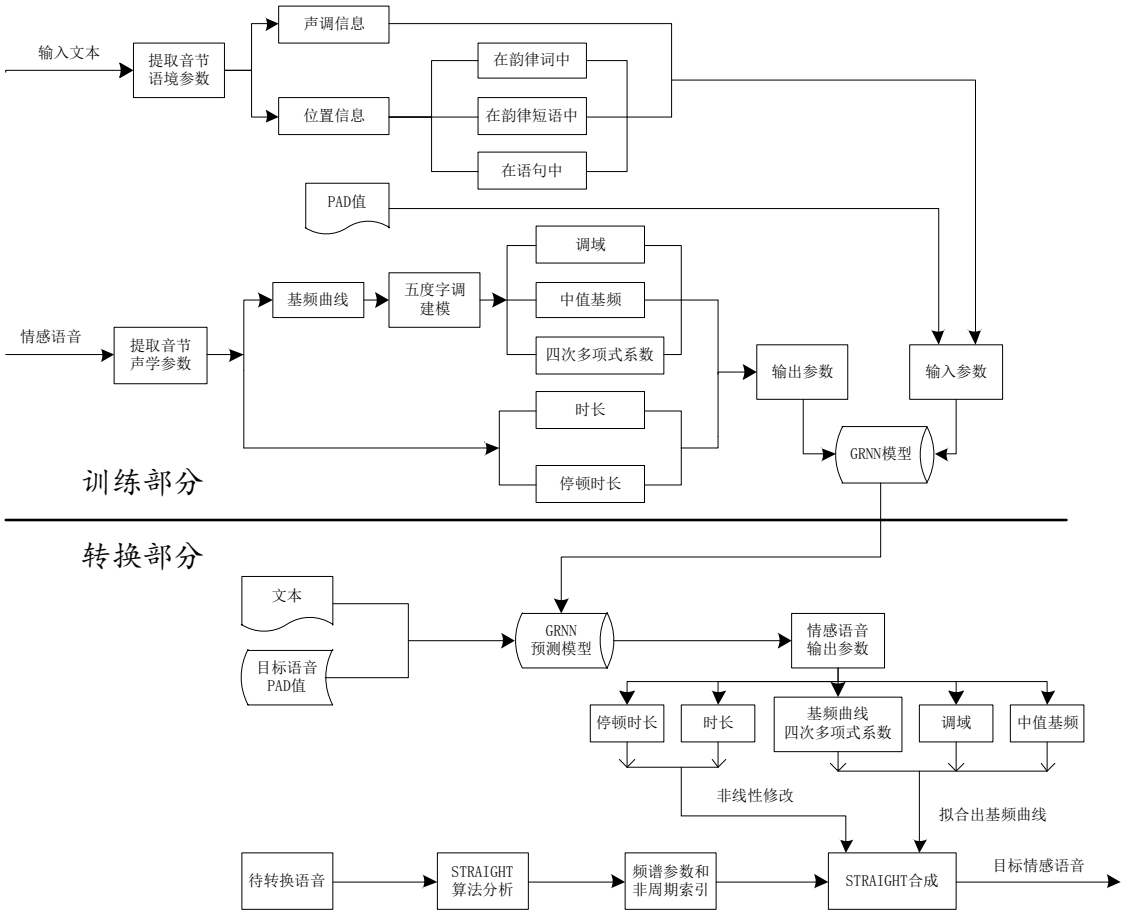


图 1.3 基于 GRNN 的韵律修改的情感语音合成方法

实验室采用基于广义回归神经网络(Generalized Regression Neural Network, GRNN)的韵律转换的方法[8]，实现了情感语音合成，如图 1.3 所示。该研究主要分

为两个阶段：训练阶段和转换阶段。在训练过程中，一方面输入训练语音文本，并提供相应的音节语境参数，另一方面，提供训练语音的情感语音文件，并提取相应的声学参数信息，并训练 GRNN 转换模型；转换阶段，一方面输入待转换语音的文本信息，及相应的文本标注信息，在 GRNN 模型库指导下得到相应的预测信息，另一方面，输入待转换语音，进行 STRAIGHT(Speech Transformation and Representation based on Adaptive Interpolation of weighted spectrogram)参数分析，根据相应的预测信息，通过 STRAIGHT 语音合成器转换得到目标情感语音。

1.2.3 基于统计参数的情感语音合成

前面已经提到，波形拼接情感语音合成系统在合成的过程中需要提供一个大型的情感语料库，但情感语音数据库的搭建工作量大，对时间和精力的投入比较高，最终合成的语音效果与构建的语料库的好坏有很大的关系，受环境影响较大，鲁棒性不高。虽然基频、时长、音强等韵律特征参数对合成语音的情感表达起着重要的作用，但音质、发声器官等声学参数，与情感语音合成特征体现也有密切的联系。

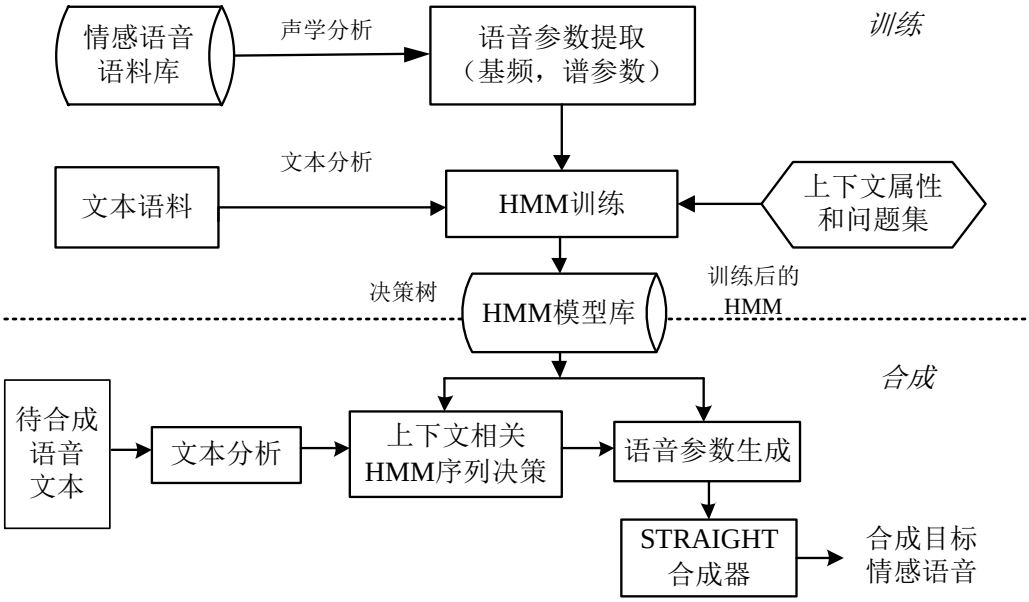


图 1.4 统计参数语音合成流程图

随着统计参数的语音合成方法[13]日益成熟，这种方法逐渐应用到情感语音合成中，它能够自动训练出情感语音的声学模型。参数化的情感语音合成方法的基本思想是，通过对输入的训练语音进行参数分解，然后对之进行声学参数建模，并构建参数化训练模型，生成训练模型库，最后在模型库的指导下，预测待合成文本的语

音参数，将参数输入参数合成器，最终得到目标情感语音。其中，基于隐 Markov 模型的统计参数语音合成方法[14-16]应用最为广泛，其流程如图 1.4 所示。

HMM 是一种统计学习模型，主要用来对半平稳随机过程进行建模。在语音信号处理中，早期利用 HMM 进行语音识别。上世纪 90 年代中期，HMM 也被引入语音合成领域。在语音合成中，早期对 HMM 训练语音合成声学模型的算法以及声学参数生成算法并不成熟，使得该方法合成的语音音质不高，远不如前面提到的基于大语料库的语音合成系统，这限制了该方法的发展和应用。经过科研人员和相关学者的不断研究，模型训练算法和参数生成算法得到改进和完善，STRAIGHT 算法的提出[17]使提取的参数更为准确，且参数合成器的性能也有了提高，从而提高了基于 HMM 的统计参数语音合成方法合成语音的质量，并被越来越多的学者应用和研究，成为当前语音合成研究的热点算法。相比基于大语料库的拼接合成方法，基于 HMM 的统计参数语音合成方法，在训练和合成的过程中，基本上不需要人工干预，对语料库数据需求也相对较少，因此这种方法要比基于大语料库的语音合成方法有一定的优势，其合成的语音音质较稳定，受语料库中训练说话人的影响较小。

1.3 论文研究内容和结构安排

1.3.1 论文研究内容

本文建立了一个多说话人多种情感的语音数据库，并针对汉语统计参数语音合成中的上下文相关标注生成，设计了包含 6 层上下文相关的标注格式，在此基础上，采用多说话人的情感语音数据，基于说话人自适应训练的统计参数语音合成方法，实现了情感语音合成。论文的主要工作和创新如下：

1.建立了一个多说话人的多种情感的语料库。在专业录音棚中，采用诱发方式激发录音人的情感，并进行录音。录制了 7 个男性说话人和 7 个女性说话人的 11 种典型情感的情感语音数据，并以 Microsoft WAV 格式（单通道、16bit、16KHz 采样频率）进行保存。

2.实现了一种面向普通话统计参数语音合成的标注生成算法。针对汉语统计参数语音合成中上下文相关标注的生成，设计了一套包含 6 层上下文相关信息的标注格式。以声韵母做为语音合成的合成基元，利用基于隐 Markov 模型(Hidden Markov Model, HMM)的统计参数语音合成方法，通过对合成语音音质的主、客观评测，验证了不同上下文信息对合成语音音质的影响。实验结果表明，本文设计的上下文相关的 6 层标注格式能够满足情感语音合成的需求。

3.提出了一种利用多个说话人的多种情感训练语料，利用统计参数语音合成方实现情感语音合成的方法。首先利用多个说话人的情感语音语料，通过说话人自适

应训练(Speaker Adaptation Training, SAT)得到多个说话人情感语音的平均音模型，然后利用目标说话人的目标情感的训练语料，经过说话人自适应变换，得到目标说话人目标情感的声学模型，进而合成出目标说话人的目标情感语音。实验结果表明，本方法合成得到的情感语音具有较高的自然度和情感相似度。

1.3.2 论文结构安排

本论文主要研究框架如图 1.5 所示，主要包含三个部分的内容：上下文相关的标注格式设计、多说话人多情感语音数据库设计和搭建、情感语音合成。

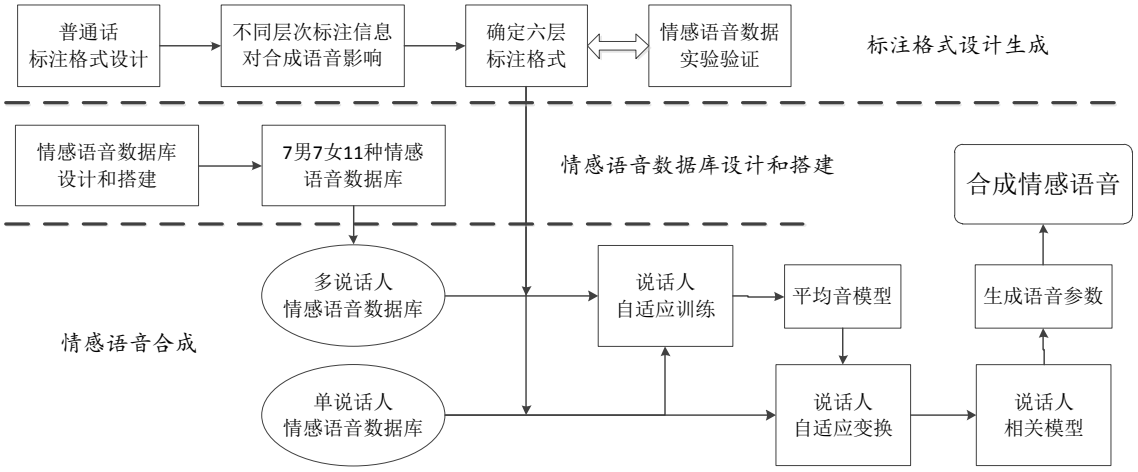


图 1.5 论文研究框架

首先，在上下文相关标注格式设计过程中，本文采用包含不同语境信息的标注格式设计方法，对普通话文本进行文本分析，生成包含不同语境信息的标注文件，并采用基于 HMM 的统计参数语音合成系统，合成出携带不同标注信息的语音，分别采用主客观评测的方法，对合成语音进行评测对比分析，验证了不同格式的标注信息对合成语音音质的影响，确定了面向普通话的统计参数语音合成的标注格式。同时，本文采用情感语音数据，对该标注格式在情感语音合成系统的可行性进行实验验证。

然后，为满足后期情感语音合成实验数据需求，本文搭建了一套多个说话人录制有多种情感的语音数据库。本文分别通过情感需求设计、情感语料获取、情感语音录制、情感语料分类等过程，搭建了 7 男 7 女包含 11 种典型情感的语音数据库。

最后，本文提出了基于多个情感说话人的说话人自适应训练的情感语音统计参数合成方法。利用多说话人情感语音数据库作为训练文件，通过说话人自适应训练得到多个说话人情感语音的平均音模型，再给定目标说话人的情感语音数据，经过说话人自适应变换，得到目标说话人目标情感的语音声学模型，并通过参数生成过程，最后合成得到目标说话人的情感语音。

论文包括六个部分的内容，如下：

第一章，前言。对本文主要研究内容——情感语音合成概述，并介绍其研究背景和现状。另外，本章还介绍了目前主要的几种情感语音合成方法，对本文情感语音合成研究方法的选择提供了理论参考。最后，概况了本论文的主要研究内容及论文结构安排。

第二章，情感语料库的设计及搭建。本章首先介绍了情感语料库构建的基本内容，并从情感分类方法、情感获取方式、文本语料设计、语音采集工具四个方面，详细的介绍了情感语料库的搭建过程。

第三章，普通话语境信息的标注生成设计。针对汉语统计参数语音合成中的上下文相关标注生成，设计了以声韵母为基元的六层上下文相关的标注格式。本章详细介绍了语音标注设计过程，实现了面向 HTS 的普通话文本标注生成程序，得到面向 HTS 系统的单音素标注文件和上下文相关的标注文件。

第四章，基于统计参数的情感语音合成。本章介绍了基于 HMM 的统计参数语音合成原理和系统搭建，并简要分析基于 HMM 的统计参数语音合成方法的特点；然后，基于 HMM 的统计参数情感语音合成系统，通过对多情感说话人语音数据自适应的方法进行模型训练，实现情感语音合成。

第五章，实验及评测。本章主要包含两个内容的实验及评测。实验一，针对汉语统计参数语音合成中的上下文相关标注生成，设计了以声韵母为基元的六层上下文相关的标注格式。搭建了基于 HMM 的统计参数语音合成系统，并通过主、客观实验评测了不同标注信息对合成语音音质的影响。实验二，分别采用单情感说话人语音数据进行自适应训练的统计参数语音合成方法和多情感说话人语音数据进行自适应训练的统计参数语音合成方法得到情感语音，并通过 MOS 和 EMOS 评测方法对两种训练模型得到的情感语音进行自然度和情感相似度对比分析。

第六章，总结与展望。对本论文主要研究工作及相应研究成果进行总结，并对下一步的研究工作进行展望。

第 2 章 情感语料库的设计及搭建

2.1 情感语料库概述

情感作为人类心理活动的组织者，是人类对客观事物所传达出的态度表现，是人际沟通的重要手段。随着智能语音技术研究的快速发展，以情感为核心的语音信号处理技术逐渐成为国内外人机智能交互领域的研究热点。同时，作为语音信号处理必不可少的一部分，研究人员对情感语料库的构建与研究[18]从未终止。

在情感语音识别中，首先要对情感语料库中的语音数据进行情感特征提取，并构建情感语音模型，再对待识别语音进行情感分析，最终确定语音内容及其情感类别；同样，在情感语音合成中，也要对情感语料库中的语音数据进行情感特征分析，并构建情感语音模型，合成语句的音素单元及情感类别的训练均来源于原始情感语音数据库，再输入待合成语音文本，通过情感语音模型获取对应基元信息，最终得到目标情感的合成语音。可以看出，情感语音数据库的质量会直接影响后期情感语音信号处理效果，所以，建立合理情感分类标准、构建高质量的情感语料库，成为情感语音研究领域的关键问题。

国内外很多研究机构对情感语料库构建的研究逐渐展开，其中，上世纪六十年代，就开始对布朗语料库(Brown-Corpus)的构建[19]；斯洛文尼亚大学分别采用英语、斯洛文尼亚语、西班牙语、法语，构建了包含八种情感色彩的 Maribor 数据库[20]；Cowie R、Cowie E 构建了包含生气、悲伤、高兴、恐惧、中性五种典型情感的 Belfast 英语数据库[21]；柏林工业大学采用德语，构建了七种情感语音的柏林 EMO-DB 数据库[22]。上世纪八十年代，国内相关科研机构，也开始着手对汉语情感语音库的构建[23]，如中科院自动化所包含高兴、生气、悲哀、惊吓、中性五种典型情感的 CASLA 数据库；清华大学和中科院心理研究所包含高兴、生气、悲伤、恐惧、中性五种典型情感的 ACCorpus 数据库。语音情感表达与说话人相关，语料库的质量会对情感色彩、语音音质、情感表达产生重要影响。

本文根据情感语音合成研究工作的需要，建立了 7 个男性说话人和 7 个女性说话人的情感语音数据，其中每个说话人的语音数据中分别包含 11 种典型情感语音，如图 2.1 所示。

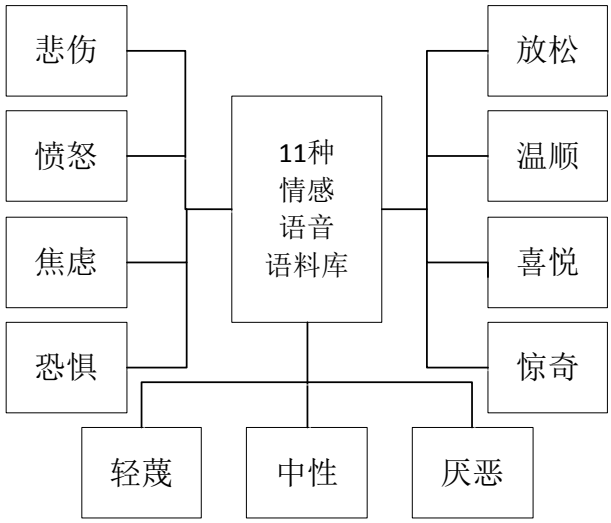


图 2.1 11 种情感语音语料库情感分类

2.2 情感语料库的构建

在情感语音合成的过程中，合成语音音质和情感语料库有直接的联系，情感语音数据库的质量直接影响后期情感语音合成效果。如何构建符合实验要求和系统需求的情感语音数据库，这涉及到情感状态的选择、情感语料获取方式、情感语音文本设计、情感数据采集方法等诸多方面的内容。如图 2.2 所示为情感语料库的基本构建流程。

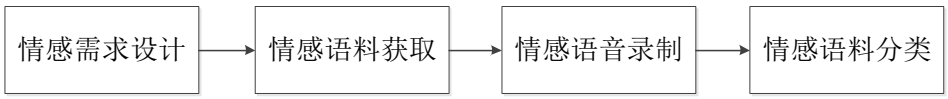


图 2.2 情感语料库的构建流程

所以说，构建一个情感分类合理、情感色彩清晰的标准情感语音数据库，是情感语音合成的重要基础，是情感语音合成系统研究工作的首要问题。同时，情感语料库也会根据搭建过程中，对情感划分、激发情感方式、文本语言设计等方面的选择，分为不同的类型[24]，如图 2.3 所示。

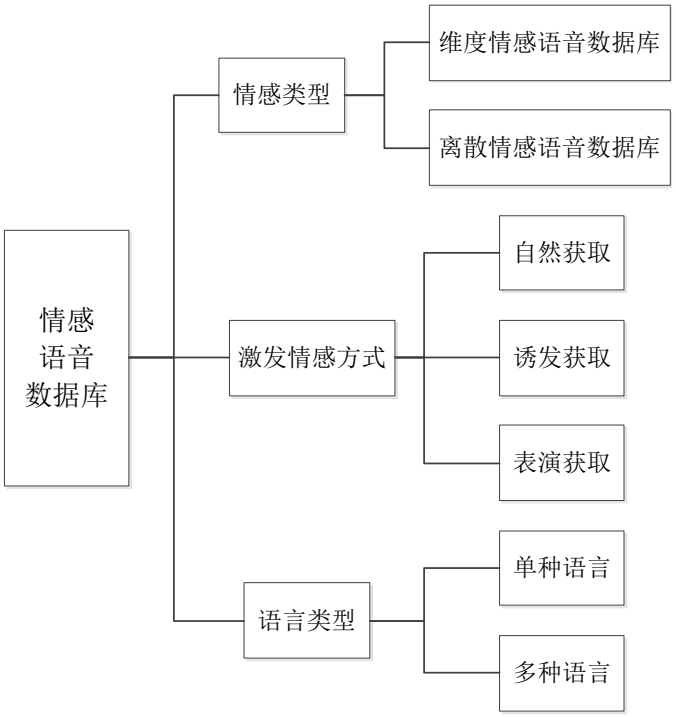


图 2.3 情感语料库的分类

2.2.1 情感分类方法

在对情感语料库构建之前，首先要对情感语音进行合理化的分类，只有这样才能对情感语音进行有针对性的收集和分析。人们普遍具有四种基本情感——喜、怒、哀、乐，在情感划分的过程中，不同情感的语音又或多或少的产生着交集，因此，对于情感的划分，往往是一种仁者见仁、智者见智的过程，至今没有统一的划分标准[25,26]。所以不同的科研机构往往根据特定的研究目的来决定情感数据的分类方法。目前，国内外研究人员对情感的划分主要有离散空间理论和维度空间理论两种方式[27]。

表 2.1 离散情感理论示例

离散情感理论	概述	举例
基本情感论	包括基本情感和派生复合情感[28]，认为情感在发生时有原始形式，每种情感有其独特的特性和模式，不同形式的情感组合组成人类的所有情感[29]；	如：Plutchik 的八种基本情感分类方法[32]，分别为恐惧、愤怒、悲伤、高兴、厌恶、惊奇、容忍和期盼；Tomkin 的八种基本情感分类方法，分别为恐惧、愤怒、痛苦、高兴、厌恶、惊奇、关心、羞愧；
三级分类模型	采用标签法把情感分为离散的形式[30]；	如：Fox[31]等人提出的三级情感分类模型；

离散情感理论认为，人类只存在几种相互独立的基本情感，根据研究目的不同，划分标准的差异，每个研究机构都会采用不同的划分方法，如表 2.1 所示。

维度空间理论认为情感是连续的，不是离散的，情感空间可以划分为几个维度，每种情感都可以映射于多维空间，都是连续体的一部分，各情感之间是可以平滑过渡的，如 Plutchik 等人提出的“情感轮(Emotion Wheel)”模型[32]。在该模型中，任意语句的情感状态都可以用一维情感矢量 E 来表示，其中，幅度值表示强度，角度表示情感方向，如图 2.4 所示。

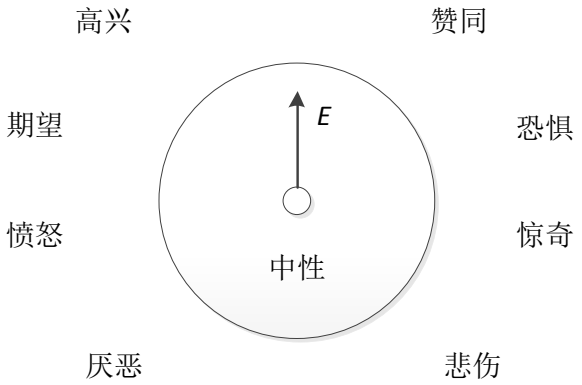


图 2.4 情感轮模型

2.2.2 情感获取方式

对情感进行明确划分，并确定情感目标后，我们就要对情感激发方法进行设计，来获取最接近原始情感的语音数据，这也是在建立情感语料库过程中，对音质效果影响最大的一个环节。目前，情感语音数据获取方法主要有三种，如表 2.2 所示，即自然情感获取、表演情感获取、诱发情感获取。

表 2.2 三种情感语音获取方式

获取方式	采集方法	特点分析	难易程度
自然获取	在预先不告知说话人的情况下，采集自然情感语音，之后再进行人工筛选	工作量较大，录音分类不确定，噪音较多，自然度较高	较难
表演获取	邀请专业人士通过表演获取	工作量较小，可操作性较高，由于录音者表演过于夸张，会导致情感和现实有所偏离	容易
诱发获取	通过环境（如看电影、阅览图片、阅读材料）刺激录音者，然后录音	工作量中等，录音前需要诱发情感，不确定录音者情感是否得到诱发	容易

从表 2.2 中可以得到如下结论，通过自然获取得到的情感语音数据，虽然情感色彩自然度较高、真实，但是获取的困难较大，得到的情感语音数据还会掺杂噪音，并且可能还会牵扯到版权等法律相关的问题；通过表演获取方式得到情感语音，虽然工作量较小，方法简单，但是采用这种方式得到的情感语音对说话人依赖较高，可能由于说话人某种情感因素过分夸大，导致情感语音数据和现实有所偏离；采用诱导方法获取情感语音，方法也相对简单，成本较低，得到的情感语音数据也较为真实。综上所述，本文采用诱导获取的方式对情感语音数据进行采集。

2.2.3 文本语料设计

在进行声学特征分析的过程中，为了体现文本相同的情况下，情感语音与中性语音的相对变化，进而得出定性结论，并进行情感语音合成方面的系统研究，设计合理的文本语料起着不容忽视的作用[30]。通过对不同情感的声学特征分析发现，文本语料的设计往往遵循以下几个准则：

1) 文本语料要有利于人体的情感激发和保持，能够反映人的内心需求和感受；

2) 文本语料要尽可能反映人们日常生活中的情感类型，能尽可能实现语音数据与选取目标情感的一致性；

3) 尽量包含语音、表情等多重信息，兼顾可能产生的语音、心理要素；

4) 对文本语料的长度进行合理设计，有利于情感表达的延续性，一般为 3~15 字为宜。

在对情感语料的文本设计上，如图 2.5 所示，本文主要选取文学小说、新闻事件、影视对白、杂志期刊等内容。

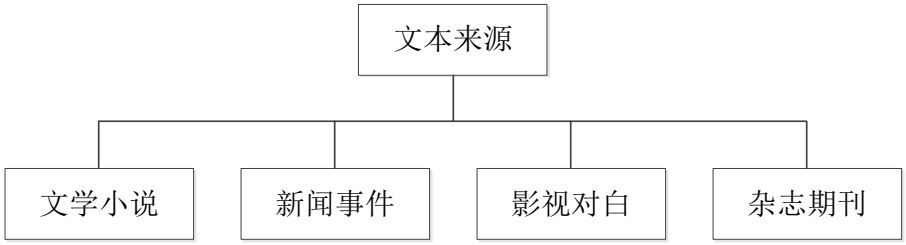


图 2.5 文本语料设计来源

2.2.4 语音采集工具

为了保证录制语音的质量，在对文本语料进行录制前，要选择符合录音条件的录音人进行录制，筛选条件主要从以下几个方面考虑。

首先，录音人年龄选择在 20-26 之间，这个年龄段的录音人一般接受过良好的语言教育，普通话发音标准，情绪表达清晰；其次，录音人的录音时间对于后期录音及后期处理有重要的影响，过长时间的录音会导致录音人情感弱化，降低情感真

实度；最后，发音人的声音应饱满、情感真实、吐字清晰。综合以上几点，最终选择了 14 名大学生作为发音人，其中 7 名为男性，7 名为女性。

如图 2.6 所示，为情感语音录影棚硬件系统框图。实验中采用了 Pro Tools 软件进行语音的采集，如图 2.7 所示，相比于其他音频处理软件，Pro Tools 处理过的音频在音质上没有损失。录音在专业录音棚中进行，硬件采用高保真话筒、高性能计算机等与 Pro Tools 配套的专业录音设备。由于纸质的发音材料在翻动过程中容易产生噪声，且发音人容易出现串行的现象，因此实验室在录音过程中，采用录音室中的计算机屏幕显示系统来解决以上问题。

本文共录制 7 男 7 女，11 种典型情感语音数据，其中每种情感语音录制 30 句，共计 4620 句(30×11×14)语料，录制得到的情感语音数据都以 Microsoft WAV 格式（单通道、16bit、16KHz 采样频率）进行保存。最后，我们还会对语音文件进行检查，对录制的语音数据进行校验和补录，以此纠正录音过程中出现的错误。

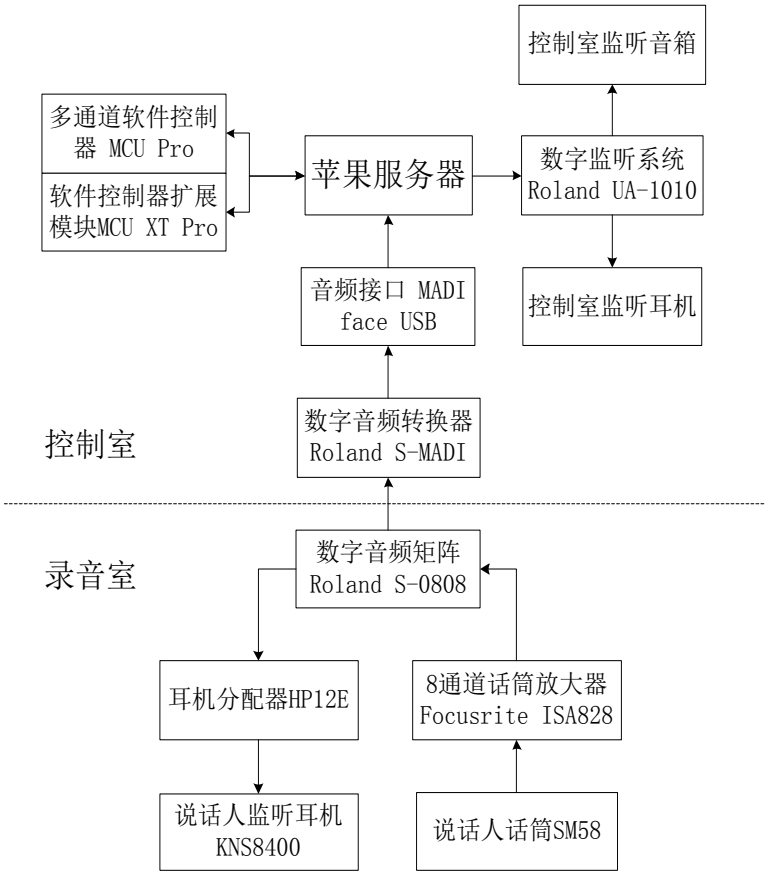


图 2.6 音频工作站系统框图

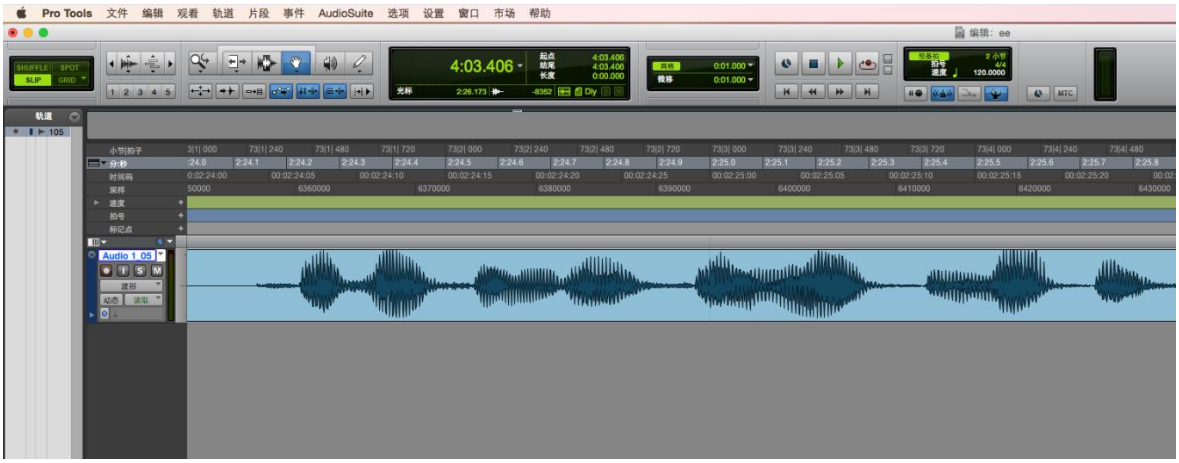


图 2.7 Pro Tools 音频和录音处理软件

2.3 本章小结

本章对情感语音和情感语音数据库进行简要概述，之后对如何进行情感语料库的构建进行了详细的阐述，主要包括情感分类、情感获取方式、文本语料的设计三个方面，最后，本章介绍了软硬件录音采集条件及情感语音采集过程，本文选择了 Pro Tools 语音采集软件，在高隔音专业录音棚中，采用专业录音系统，进行原始情感语音的采集。

第 3 章 普通话语境信息的标注生成设计

汉语统计参数语音合成系统主要包括前端的文本分析和后端的语音合成。其中，前端的文本分析主要用来获得合成文本的发音信息及其上下文相关信息；后端的语音合成用来生成文本对应的声学参数并合成出语音。前端文本分析获得的读音信息和上下文信息对合成语音的音质有至关重要的影响。

随着对自然语言处理的要求越来越高[33]，对汉语文本分析的研究也越来越广泛，主要包括句法分析[34]、自动语义标注[35,36]、韵律结构预测[37,38]、自动分词[39,40]等。另一方面，自然语言处理涉及较多的领域和分支，而且各个领域和分支都有一定的独立性，部分技术已达到或者基本达到实用化的程度，并在实际应用中发挥着巨大作用，比如语音识别、自动文摘、文字识别、搜索引擎等。但现有的研究缺少面向普通话统计参数语音合成的文本分析，也缺少上下文信息对合成语音音质影响的研究。

本章针对普通话统计参数语音合成中的上下文相关的标注生成，设计了包含 6 层上下文相关的标注格式。

3.1 汉语普通话的文本分析

文本分析是文语转换系统的前端工作，如图 3.1 所示。在文语转换中，文本分析起着最基础而又关键的作用。文语转换系统首先通过分析得到输入文本的音调、韵律、音节等信息，并生成得到相应的语境信息标注文件，这个过程称为文本分析，再通过后端的语音合成模块把这些信息转换成输出语音。在整个过程中，各项处理规则都是以文本分析为基础的，文本分析为后端语音合成提供了重要依据，文本分析的效果会直接影响到合成语音的自然度、准确性。

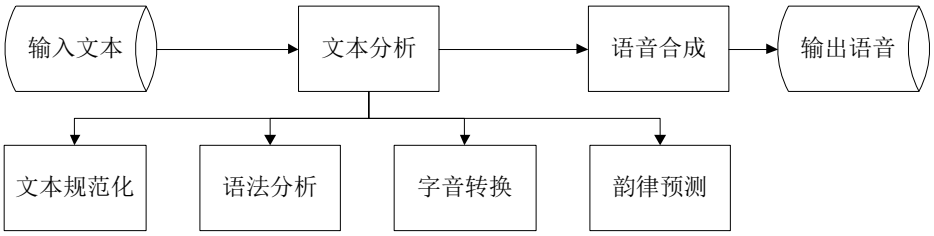


图 3.1 文语转换系统总框图

文本分析可以划分为四个模块的内容，如图 3.1 所示。本文以汉语普通话的声韵母为语音合成基元，对输入的汉语文本，借助于语法词典、语法规则库的指导，通过文本规范化、语法分析、韵律预测分析、字音转换，依次获得输入文本的语句信息、词信息、韵律结构信息和每个汉字的声韵母，从而获得输入普通话文本的语

音合成基元（声韵母）的信息以及每个语音合成基元的上下文相关信息，最终生成语音合成后端所需的单音素标注和上下文相关的标注，其过程如图 3.2 所示。

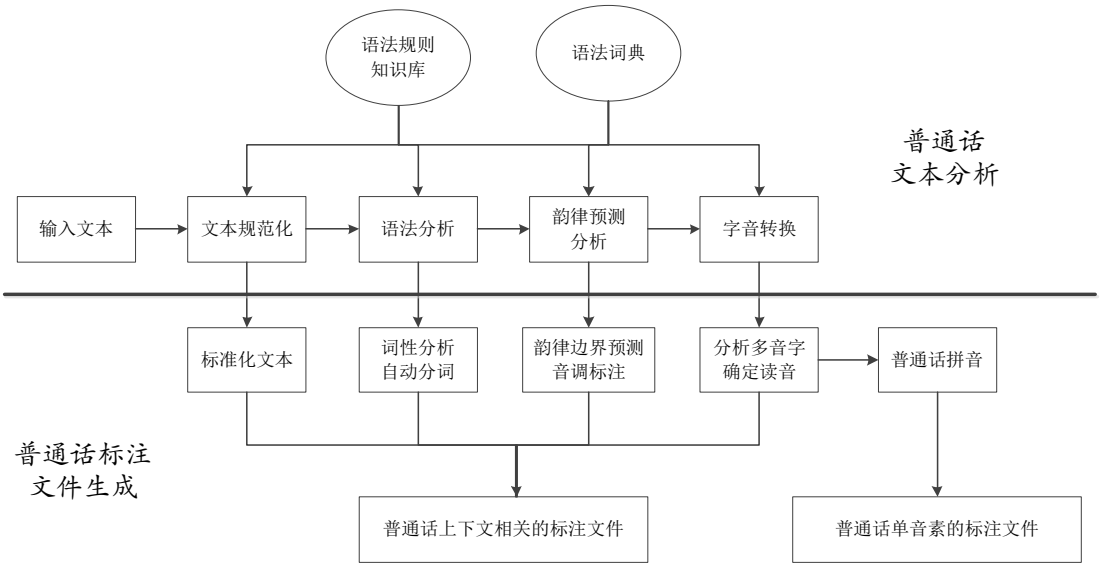


图 3.2 普通话文本分析流程图

3.1.1 文本规范化

普通话文本中除了正常的文字外，还会经常出现简略词、日期、公式、号码等文本信息，这就需要通过文本规范化，对这些文本块进行处理，否则合成出的语音就会出现语法错误、语义不完整等情况，从而导致语音合成效果的降低。另外，在不同的上下文环境下，一些非文字文本也会表达不同的含义，比如：“小明体重是 128 斤”中的“128”应该规范为“一百二十八”，而“G128 次列车”中的“128”应该规范为“幺 二 八”；还有一些不同的输入文本，却表达着相同的含义，比如：日期的多种写法，“2016-05-15”、“2016 年 5 月 15 号”、“2016/05/15”。因此，文本规范化作为普通话文本分析中不可缺少的一个模块，直接影响着普通话文本拼音信息的准确度，并最终决定输出语音的合成效果。

文本规范化[41,42]就是通过对上下文文本的分析，结合上下文环境，把输入文本中正常文本以外的非标准文本信息转化为对应文字的过程，其过程如下：

1) 结合有限状态机[43](Finite State Machine, FSM)和最大匹配原则，在词典的指导下，把文本中不标准字符串识别为非标准文本词；

2) 选取人工设计的特征模板，采用最大熵算法的统计模型，进行训练建模；

假设 A 为待消歧问题所有可能的集合，a∈A 判定歧义问题为某种可能的结果，B 为消歧点所在上下文信息的集合，b∈B 为某种歧义问题可能结果所对应的上下文

信息，从而计算判定结果 a 的条件概率 $p(a|b)$，即通过最大熵模型选取条件概率 $p(a|b)$ 最大的结果最为判定结果：

$$\hat{p}(a|b)=\arg\max_{p\in P}H(p) \tag{3.1}$$

其中，P 指的是与样本相吻合的所有概率分布的集合。

因为

$$\begin{aligned}H(p)&=H(A|B)\\&=\sum_{b\in B}p(b)H(A|B=b)\\&=-\sum_{a,b}p(b)p(a|b)\log p(a|b)\end{aligned} \tag{3.2}$$

因此，

$$\begin{aligned}\hat{p}(a|b)&=\arg\max_{p\in P}H(p)\\&=\arg\max_{p\in P}(-\sum_{a,b}\hat{p}(b)p(a|b)\log p(a|b))\end{aligned} \tag{3.3}$$

3) 在规则模型的指导下，对非标准词进行消歧，规范为标准文本。

3.1.2 语法分析

高质量的语法分析可以获得语句的语法结构，实现对语句的精确理解。语法分析主要结合词性标注、句法分析两部分的内容实现对语句的自动分词。首先需要对输入的文本段落进行划分得到语句，再根据标点符号来确定语句划分边界，但是，由于汉语文本没有严格的词边界，如何对输入文本进行自动分词又成为语法分析中一个重要环节[41]。本文采用基于词的三元语法模型[45]，结合最大匹配算法进行自动分词，过程如下：

1) 根据词典采用最大匹配方法对句子进行简单匹配，找出所有可能的词典词，并把找到的词典词和所有单个字共同组成词候选集 $C=c_1c_2\cdots c_N$；

2) 根据词典词候选集 C 中所有单元进行词性标注，并构造三元文法切分词图，其中概率 P(C)为不同切分方法的匹配代价：

$$P(C)=P(c_1)P(c_2|c_1)\prod_{i=3}^{N}P(c_i|c_{i-2}c_{i-1}) \tag{3.4}$$

3) 采用 Viterbi 搜索算法，选取代价最小的切分方法作为最后的分词结果。

3.1.3 字音转换

字音转换，也就是确定每个字的读音，把文本中的文字转换为音节的读音的过程。字音转换中的难点是多音字读音的确定。本文采用基于规则的多音字自动注音方法[46]确定多音字的读音。过程如图 3.3 所示。

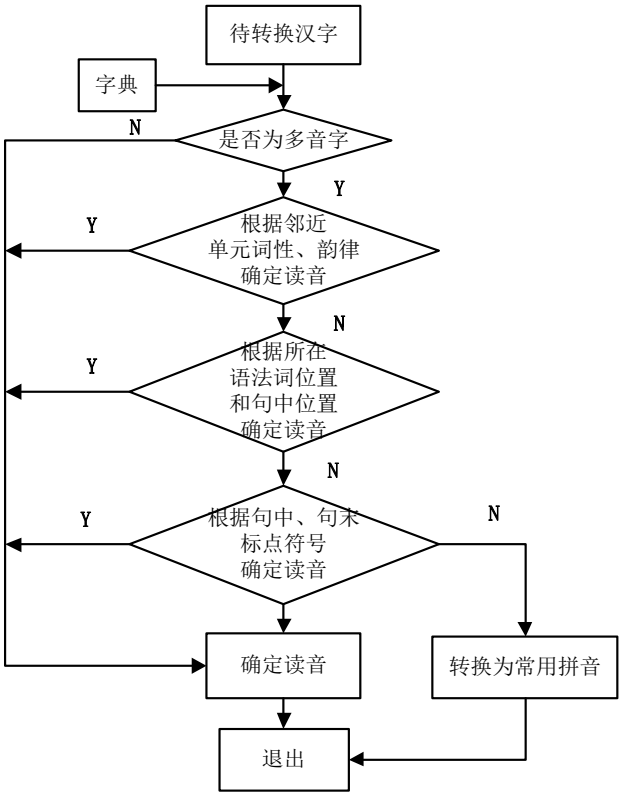


图 3.3 字音转换自动标注流程图

首先，在字典的指导下，对句中文字进行分类，如果该字只有一个读音，则转换为标准拼音；如果该字有两个或两个以上读音，则参照相邻单元词性、词长等特征信息确定读音。

3.1.4 韵律预测分析

文语转换系统中，准确的韵律预测分析对提高合成语音的流畅度、清晰度、自然度，起着最重要而又关键的作用，并最终可以有效提高合成语音的节奏感、真实感。根据语音学的研究，韵律特征具有层级结构，如图 3.4 所示，汉语普通话的韵律层级结构，一般可分为四层[47]。其中，韵律词层、韵律短语层的预测对合成语音的自然度的影响最大[38]。

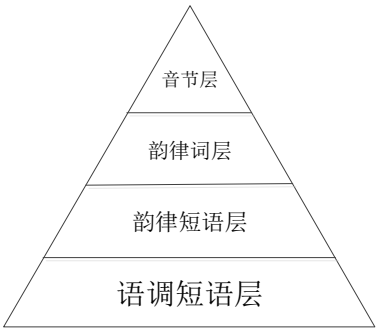


图 3.4 汉语普通话的韵律层级结构

为了提高韵律预测分析的精度，本文利用语法词的属性（如词性、词长）、语法树的结构信息、语法词在语法树中所处层级等信息作为输入特征，采用TBL(Transformation-based error driven learning algorithm)算法[47,48]，实现对文本的韵律边界预测。TBL 算法分为训练和预测两个步骤，在训练步骤，由人工设计一套规则模板，如果在训练的过程中，遇到和预定义的规则模板不匹配的情况，TBL 算法会进入预测步骤，从实例中自动学习新规则，并对原有规则进行添加、修改、转换。重复上述过程，最终训练得到得分最高的训练规则集，作为理想的实例化规则。

3.2 上下文相关标注格式

3.2.1 标注格式设计内容

对于标注格式的设计，本文设计一种以普通话声韵母为语音合成基元的上下文相关的标注格式，其中这些基元还要包括静音段和停顿段，如表 3.1 所示。

表 3.1 静音、停顿符号表

类型	定义值
起始或结尾的静音	sil
句子中的短停	pau

如图 3.5 所示，为普通话孤立音节的五种音调[49]，音素是构成韵律词的基本单元，对于韵律词边界的的描述主要依靠声调的变换，或者句子中的音变引起的声调变换等[56]。

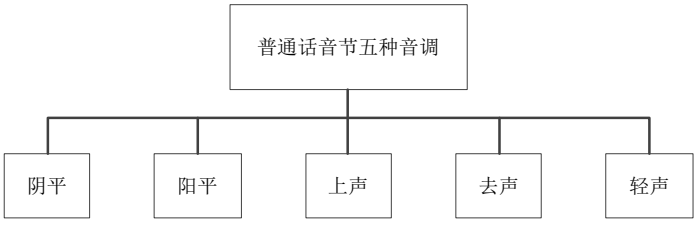


图 3.5 普通话音节五种音调

如表 3.2 所示，为韵律词层的声调标注表示方法：

表 3.2 声调符号表

类型	符号表示
声调	m1 m2 m3 m4 m5
文本分析中的变音	m1 m2 m3 m4 m5 ms1 ms2 ms3 ms4

词层用来表示当前词的类型，其中根据不同的词类型，需要设定不同的表示方法，普通话词性表示见附录 A。

句子层主要表示句子的类型，一般陈述句用 d 表示，疑问句用 q 表示，祈使句用 i 表示，感叹句用 e 表示，其中陈述句会比较多，如表 3.3 所示。

表 3.3 句子类型表

句子类型	符号
陈述句	d
疑问句	q
祈使句	i
感叹句	e

3.2.2 上下文相关标注格式

在标注格式的设计中，需要考虑发音基元及其前后基元的信息，以及发音基元所在各标注层的语境相关信息，因此，本文设计了一套包含 6 层的上下文相关的标注格式，如表 3.4 所示。

表 3.4 上下文相关的标注格式

声韵母层	p1^p2-p3+p4=p5@p6_p7
音节层	/A:a1_a2-a3_a4#a5 /B:b1_b2!b3_b4#b5@b6!b7+b8@b9#b10_b11 /C:c1+c2-c3 =c4 #c5
韵律词层	/D:d1-d2 /E:e1&e2^e3_e4 /F: f1-f2
韵律短语层	/G:g1-g2 /H:h1-h2@h3+h4 /I:i1-i2
词组层	/J: j1^j2= j3-j4 /K:k1=k2_k3^k4&k5_k6 /L:l1^l2=l3-l4
语句层	/M:m1#m2+m3+m4!m5

声韵母层中，p1-p7 代表发音基元信息。其中，p1-p5 分别表示前前基元、前一基元、当前基元、后一基元、后后基元；p6、p7 分别从前向、后向记录当前音素在音节中的位置信息。本文选取了声韵母作为语音合成的基元，包括 21 个声母、39

个韵母、静音段、停顿段。其中，语句起始段、结尾段的静音和句中短停分别定义为 sil 和 pau。

音节层分为 A、B、C 三个部分，分别代表前一音节(A)、当前音节(B)和后一音节(C)的相关信息。其中，a1-a2、b1-b2、c1-c2 分别代表各音节的首、末基元。声调是合成语音自然度、真实感的重要体现，利用 a3-a4、b3-b4、c3-c4 分别代表各音节在词典和文本分析中的声调类型，普通话的孤立音节五种音调，如图 3.5 所示。a5、b5、c5 分别代表各音节包含的音素个数。b6-b11 分别从前向、后向记录当前音节在词、韵律词、韵律短语中的位置信息。

词层包含 D、E、F 三个部分，分别代表前一词(D)、当前词(E)和后一词(F)的相关信息。其中，d1、e1、f1 表示各个词的词性；d2、e2、f2 代表各词中包含的音节个数；e3-e4 分别从前向、后向记录当前词在韵律词中的位置信息。

韵律词层包含 G、H、I 三个部分，分别代表前一韵律词(G)、当前韵律词(H)、后一韵律词(I)的相关信息。其中，g1-g2、h1-h2、i1-i2 分别代表各个韵律词中包含的音节个数和词个数；h3-h4 分别从前向、后向记录当前韵律词在韵律短语中的位置信息。

韵律短语层包含 J、K、L 三个部分，分别代表前一韵律短语(J)、当前韵律短语(K)、后一韵律短语(L)的相关信息。其中，j1、k1、l1 分别代表各韵律短语的语调类型；j2-j4、k2-k4、l2-l4 分别代表各韵律短语中包含的音节个数、词个数和韵律词个数；k5-k6 分别从前向、后向记录当前韵律短语在当前语句中的位置信息。

M 层为语句层，其中，m1 是语句的句调类型，包括陈述句、疑问句、祈使句、感叹句 4 种类型；m2-m5 分别记录当前语句中包含的音节个数、词个数、韵律词个数、韵律短语个数。

如附录 B 所示，为上下文相关标注格式设计字母含义表，该表详细介绍了六层标注格式所包含内容，及每个字母所代表的具体含义。

3.3 标注文件生成实例

根据对上下文语境信息的标注设计格式，实现了面向基于隐 Markov 模型统计参数语音合成系统的汉语普通话上下文相关的标注生成程序，生成了基于隐 Markov 模型统计参数语音合成系统中训练过程所需的单音素标注文件和上下文相关的标注文件，生成的两种标注文件和训练语音文件一一对应。上下文属性单元标注文件也可根据实验需求设计包含时间信息的单音素标注文件，标注文件的每行信息一般会包含音素的开始时间、结束时间。

3.3.1 单音素标注文件生成实例

如图 3.6 所示为一个训练语句的单音素标注文件实例，句子为：一群活泼可爱的男孩子在草地踢足球。从图中可以看出，生成的单音素标注文件标注出了语句中的所有音素信息及其位置信息。其中，sil 表示这一时间段是静音段；pau 表示为短停段；q、vn、h 等都是拼音单元（音素）。

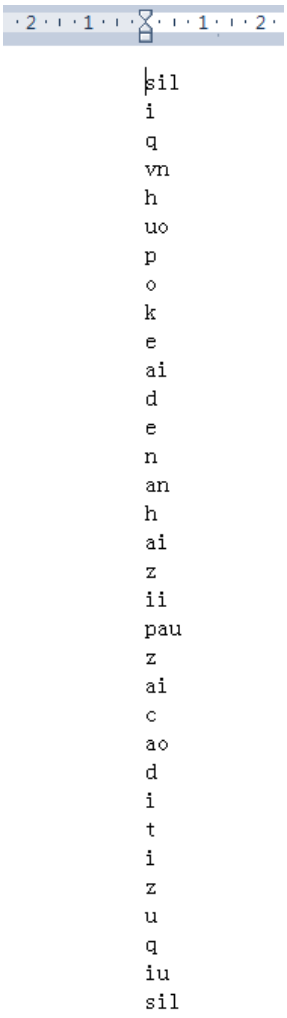


图 3.6 单音素标注文件实例

3.3.2 上下文相关标注文件生成实例

图 3.7 所示的是生成的上下文相关的标注文件实例，它显示了上下文相关的标注文件中基元音素所包含的上下文相关信息。在文本分析中，一方面经过文本规范化、字音转换生成单音素标注文件，另一方面经过文本规范化、语法分析、韵律分析分析过程，提取上下文相关信息，再结合单音素标注文件生成上下文相关的标注文件。

```
xx^xx-sil+i=q@x_x/A:xx_xx-xx_xx=xx/B:xx_xx!xx_xx#xx@xx!
xx+xx@xx#xx_xx/C:xx+i-m1=m4#1/D:xx-xx/E:xx&xx^xx_xx/F:q-2/G:xx-xx/H:xx-
xx@xx+xx/I:2-1/J:xx&xx=xx+xx/K:xx=xx_xx^xx&xx_xx/L:d^10=4-4/M:0#16+10+6!
2
xx^sil-i+q=vn@1_1/A:xx_xx-xx_xx=xx/B:xx_i!m1_m4#1@1!2+1@2#1_10/C:q+vn-m2
=m2#2/D:xx-xx/E:q&2^1_1/F:q-2/G:xx-xx/H:2-1@1+4/I:2-1/J:xx&xx=xx+xx/K:d=
10_4^4&1_2/L:d^6=2-2/M:0#16+10+6!2
sil^i-q+vn=h@1_2/A:xx_i-m1_m4=1/B:q_vn!m2_m2#2@2!1+2@1#2_9/C:h+uo-m2=m2#
2/D:xx-xx/E:q&2^1_1/F:q-2/G:xx-xx/H:2-1@1+4/I:2-1/J:xx&xx=xx+xx/K:d=10_4
^4&1_2/L:d^6=2-2/M:0#16+10+6!2
i^q-vn+h=uo@2_1/A:xx_i-m1_m4=1/B:q_vn!m2_m2#2@2!1+2@1#2_9/C:h+uo-m2=m2#
2/D:xx-xx/E:q&2^1_1/F:q-2/G:xx-xx/H:2-1@1+4/I:2-1/J:xx&xx=xx+xx/K:d=10_4
^4&1_2/L:d^6=2-2/M:0#16+10+6!2
q^vn-h+uo=p@1_2/A:q_vn-m2_m2=2/B:h_uo!m2_m2#2@1!2+1@2#3_8/C:p+o-m5=m5#
2/D:q-2/E:q&2^1_1/F:a-2/G:2-1/H:2-1@2+3/I:3-2/J:xx&xx=xx+xx/K:d=10_4^4&1
_2/L:d^6=2-2/M:0#16+10+6!2
vn^h-uo+p=o@2_1/A:q_vn-m2_m2=2/B:h_uo!m2_m2#2@1!2+1@2#3_8/C:p+o-m5=m5#
2/D:q-2/E:q&2^1_1/F:a-2/G:2-1/H:2-1@2+3/I:3-2/J:xx&xx=xx+xx/K:d=10_4^4&1
_2/L:d^6=2-2/M:0#16+10+6!2
h^uo-p+o=k@1_2/A:h_uo-m2_m2=2/B:p_o!m5_m5#2@2!1+2@1#4_7/C:k+e-m3=m3#
2/D:q-2/E:q&2^1_1/F:a-2/G:2-1/H:2-1@2+3/I:3-2/J:xx&xx=xx+xx/K:d=10_4^4&1
_2/L:d^6=2-2/M:0#16+10+6!2
uo^p-o+k=e@2_1/A:h_uo-m2_m2=2/B:p_o!m5_m5#2@2!1+2@1#4_7/C:k+e-m3=m3#
2/D:q-2/E:q&2^1_1/F:a-2/G:2-1/H:2-1@2+3/I:3-2/J:xx&xx=xx+xx/K:d=10_4^4&1
_2/L:d^6=2-2/M:0#16+10+6!2
p^o-k+e=ai@1_2/A:p_o-m5_m5=2/B:k_e!m3_m3#2@1!2+1@3#5_6/C:xx+ai-m4=m4#
1/D:q-2/E:a&2^1_2/F:a-1/G:2-1/H:3-2@3+2/I:3-1/J:xx&xx=xx+xx/K:d=10_4^4&1
_2/L:d^6=2-2/M:0#16+10+6!2
o^k-e+ai=d@2_1/A:p_o-m5_m5=2/B:k_e!m3_m3#2@1!2+1@3#5_6/C:xx+ai-m4=m4#
1/D:q-2/E:a&2^1_2/F:a-1/G:2-1/H:3-2@3+2/I:3-1/J:xx&xx=xx+xx/K:d=10_4^4&1
_2/L:d^6=2-2/M:0#16+10+6!2
k^e-ai+d=e@1_1/A:k_e-m3_m3=2/B:xx_ai!m4_m4#1@2!1+2@2#6_5/C:d+e-m5=m5#
2/D:q-2/E:a&2^1_2/F:a-1/G:2-1/H:3-2@3+2/I:3-1/J:xx&xx=xx+xx/K:d=10_4^4&1
```

图 3.7 上下文相关的标注文件实例

如图所示，第一行的 sil^i-q+vn=h 为声韵母层内容，该标注内容显示当前单元为声母/q/，它的前前单元为静音段标注 sil、前一个音素单元为韵母/i/、后一个音素单元为韵母/vn/、后后单元为声母/h/。根据附录 B 可知，/A、/B、/C 分别表示音节层的相应语境信息，/D、/E、/F 分别表示词层的相应语境信息，/G、/H、/I 分别表示韵律词层的相应语境信息，/J、/K、/L 分别表示韵律短语层的相应语境信息，/M 表示语句层的语境信息。

3.4 问题集的设计和决策树聚类

在普通话语音实际的发音过程中，会有协同发音的现象，每个发音基元的发音都会受左右基元的影响。为了在语音合成时能够根据发音基元的语境信息产生最优的声学参数，需要设计一套能够反映连续语流中每个发音基元的语境信息的表示方

法。同时，为了能够对声学模型按照语境信息聚类，也需要根据语境信息设计一套决策树聚类所需的问题集。

3.4.1 问题集的设计

在单音素模型训练好之后，我们需要对声学参数模型进行决策树聚类，从而对模型更好的进行划分，这就需要在问题集的指导下，利用决策树聚类算法来建立包含上下文语境信息的 HMM 模型。问题集的设计主要是根据上下文相关标注信息，对发音单元基本特征做出的分类，比如音素的位置信息，音素的前后单元信息等。

如图 3.8 所示，本文设计了一套面向普通话的上下文相关的问题集，该问题集包含了 3000 多个上下文相关的问题，基本覆盖了上下文相关语境信息的所有特征。格式如下：

QS 问题表达式 {答案 1，答案 2，答案 3，……}

其中，每个问题都是以 QS 命令开头，问题集的答案可以有多个，中间以逗号隔开，答案是一个包含通配符的字符串。当问题表达式为真时，该字符串成功匹配标注文件中的某一行标注，在 HTS 系统中，问题集主要和上下文相关的标注文件相匹配。

```
QS "LL==Fricative"          {f^*,s^*,sh^*,x^*,h^*,lh^*,hy^*,hh^*}
QS "LL==Front_Fricative"          {f^*}
QS "LL==Front_Stop"         {b^*,p^*}
QS "LL==Front_Vowel"              {i^*,v^*,ei^*}
QS "LL==Glottal"            {hh^*}
QS "LL==Initial"
      {b^*,ch^*,c^*,d^*,f^*,g^*,h^*,j^*,k^*,l^*,m^*,n^*,p^*,q^*,r^*,sh^*,s^*,t^
*,x^*,zh^*,z^*,lh^*,hh^*,ny^*,gy^*,ky^*,hy^*,ng^*,w^*,y^*,mb^*,nz^*,nd^*,nzh^*,
nj^*,ngy^*,ngg^*}
QS "LL==Lab_Dental"         {f^*}
QS "LL==Lateral"            {l^*}
QS "LL==Man_Final"
      {er^*,ia^*,iii^*,ua^*,ui^*,ai^*,ei^*,ao^*,ou^*,iao^*,uai^*,uei^*,ian^*,ia
ng^*,uan^*,uen^*,uang^*,ueng^*,ong^*,van^*,vn^*,iong^*}
```

图 3.8 问题集设计的部分实例

3.4.2 决策树聚类

为了提高建模的精度，我们需要根据问题集，采用决策树聚类对模型进行训练和划分，从而建立上下文相关的 HMM 模型。如图 3.9 所示，决策树是一个二叉树，每一个叶子节点都包含着一个上下文相关的问题，两个子节点分别代表该问题的答

案是否符合。叶子节点包含着状态输出分布，通过问题集使用决策树进行上下文相关的聚类，可以从上下文中获得语音单元的模型参数，从而建立音素的上下文相关模型。

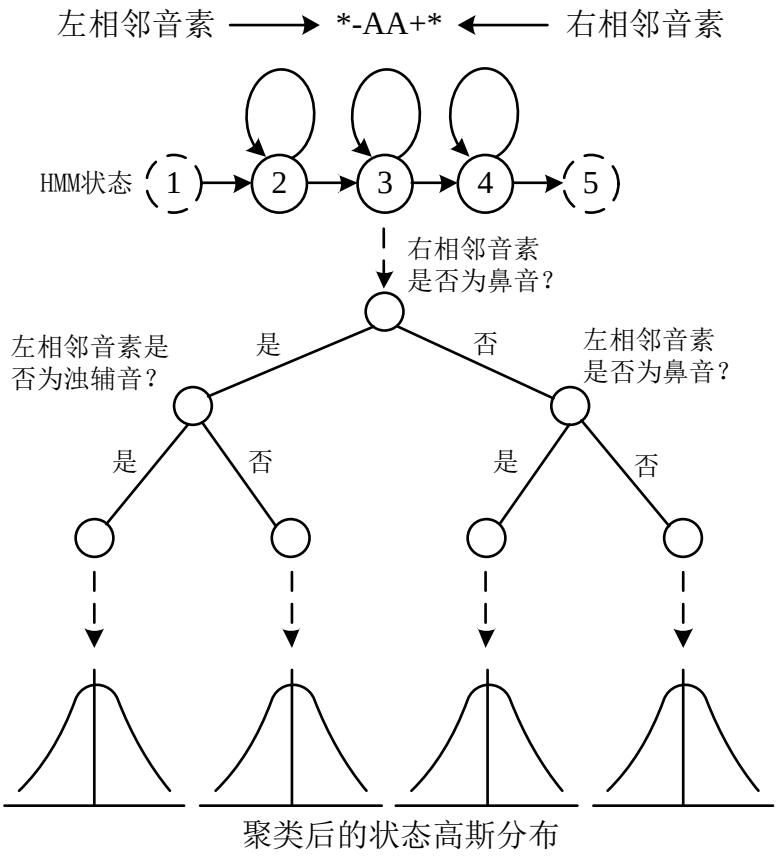


图 3.9 决策树聚类

3.5 本章小结

本章从普通话发音的基本单元出发，分析普通话中语境相关信息，并以普通话声韵母为基元音素，设计了一套面向汉语普通话的上下文相关的语境信息标注格式。通过文本规范化、语法分析、韵律预测、字音转换等文本分析过程，设计了包含六个层级的标注格式，介绍了语音的标注格式设计方法，实现了面向 HMM 统计参数语音合成的文本标注程序，生成了面向 HMM 统计参数语音合成系统的单音素标注文件和上下文相关的标注文件。

第 4 章 基于统计参数的情感语音合成

4.1 基于 HMM 统计参数的语音合成

隐 Markov 模型(Hidden Markov Model, HMM)是一种具有双重的随机过程的机器学习模型，观察到的事件是状态的随机函数，我们不知道模型所经过的状态序列，只知道状态的概率函数。20 世纪 70 年代中期，隐 Markov 模型开始应用于语音识别[50]，到 90 年代中期，隐 Markov 模型逐渐被引入语音合成上来。接下来本文对基于隐 Markov 模型的统计参数语音合成方法的基本原理进行介绍，并搭建实现了基于隐 Markov 模型的统计参数语音合成系统。

4.1.1 基于隐 Markov 模型的统计参数语音合成系统

本文采用基于隐 Markov 模型的统计参数语音合成系统进行语音合成，该系统可分为训练阶段和合成阶段[51]，基本框图如图 4.1 所示。

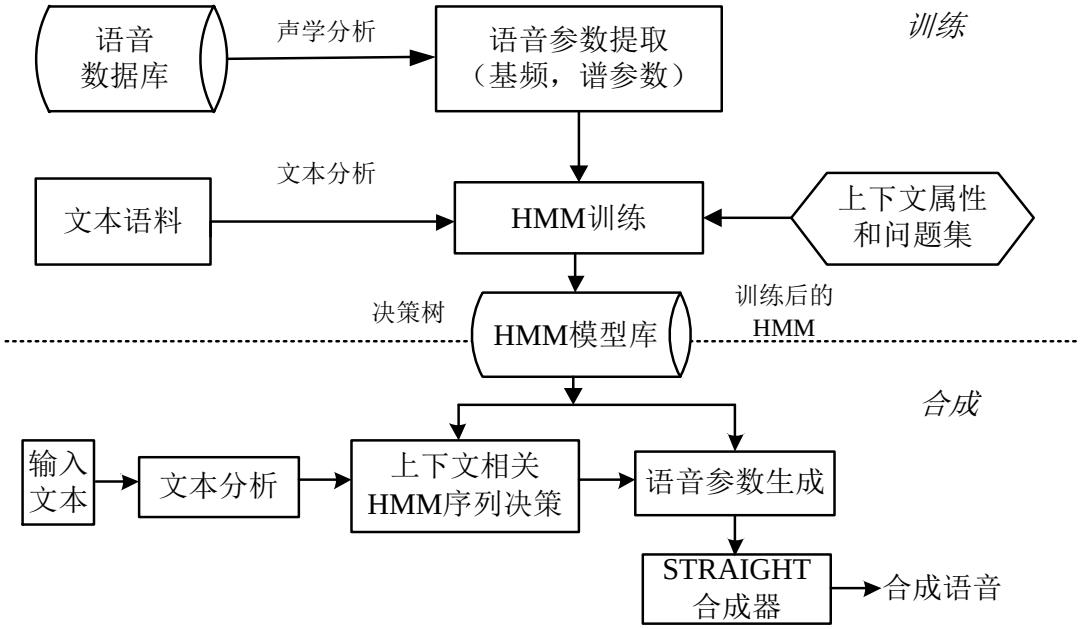


图 4.1 基于隐 Markov 模型的统计参数语音合成系统

首先，我们可以把训练过程分为预处理阶段和隐 Markov 模型训练阶段。在预处理部分，一方面对文本语料进行文本分析，得到训练语音文本的单音素标注文件和上下文相关信息标注文件，其中上下文相关信息标注文件包含了对语音参数建模会产生影响的上下文相关信息，比如前后单元声调信息、单元位置信息、韵律信息、分词信息等。另一方面通过对训练语音数据库进行声学参数提取分析，主要为基频和谱参数，然后对声学参数分析建模，在汉语发音过程中，由于清音和无音段没有

基频，一般不会采用连续概率分布隐 Markov 模型进行基频建模，本文采用多空间概率分布 HMM(Multi-space probability distribution, MSD-HMM)实现基频建模。

在隐 Markov 模型训练过程中，会出现因为训练数据过少，而使模型出现过拟合现象，本系统根据语言学、语音学等相关方面的知识来选择一些对声学参数会产生影响的上下文属性，来提高训练模型的鲁棒性，设计了用于决策树聚类的问题集。然后训练每个发音基元的 HMM 模型，在问题集的指导下，利用决策树对模型进行聚类，得到语音合成所需的 HMM 模型库。

在合成阶段，首先通过文本分析得到待合成文本的每个发音基元的上下文相关的标注文件；然后根据每个基元的上下文相关标注信息，利用决策树从 HMM 模型库中挑选出发音基元的 HMM 模型，并拼接到一起生成语句 HMM 模型；最后采用参数生成算法得到待合成语句的声学参数，并利用 STRAIGHT 算法合成出目标语音。

4.1.2 基于隐 Markov 模型的参数语音合成方法的特点

基于隐 Markov 模型的统计参数语音合成方法[52]和基于波形拼接的合成方法有明显的不同。基于隐 Markov 模型的统计参数语音合成方法具有鲜明的技术特点：

1）系统可实现快速、自动化构建。基于隐 Markov 模型统计参数的语音合成方法可以在较少人工干预模型的情况下，自动实现声学模型训练、参数预测、语音合成等过程，在较短的时间内实现系统构建。虽然 HMM 模型在自动训练的过程中会耗费较多时间，但这问题会随着硬件水平的提高得到改善。而采用波形拼接的语音合成方法，大部分需要进行人工调整，耗时耗力。

2）合成语音平滑、韵律流畅。本系统得到的合成语音平滑，韵律流畅，对于不同文本的适应性较强，鲁棒性高。采用基于单元挑选与波形拼接的语音合成方法，拼接处不连续和基频不稳定的现象时有发生，同时对语音训练数据依赖性较高。

3）语音数据量较少。针对波形拼接的语音合成系统的训练语音的录制，往往需要很长时间。而采用统计参数的语音合成系统，语音数据量要求较少，一般 1 个小时的语音库数据量即可满足系统训练需求，有效的降低了系统构建成本。

4）系统灵活度高。基于隐 Markov 模型的统计参数语音合成方法，仅需提供目标说话人较少的语音数据，即可采用说话人自适应训练的方法，实现目标说话人的语音合成，系统灵活度较高。而采用基于大语料库的拼接合成的方法，当合成语音的音色或风格发生改变时，研究人员就要对整个系统语音数据进行重新采集，应用领域具有一定的局限性。

5）系统存储容量小。采用基于隐 Markov 模型的统计参数合成，在合成的过程中，只需要提供训练得到的声学参数模型，不需提供原始语音，有效的降低了系统

存储空间，可以方便的实现在嵌入式系统设备的移植。而拼接合成方法，需要大量的原始语音，系统需求较大。

4.2 基于多情感说话人自适应的情感语音合成

采用基于隐 Markov 模型的统计参数语音合成方法实现高质量的情感语音合成，对情感语料库的采集工作量要求较大，合成的情感语音音质不高，很难直接把这种方法应用到情感语音合成中来。所以，本文在情感语音合成系统训练过程中，采用多个情感说话人同一情感语音数据进行自适应训练[5,6]，得到多个情感说话人语音数据的平均音模型，这样可以减少由于说话人情感表达的差异所造成的影响，进而提高合成情感语音的自然度和情感相似度。

得到多个说话人情感语音平均音模型后，我们只需给定少量目标说话人待合成情感语音，通过说话人自适应变换过程，得到目标说话人自适应情感语音数据模型库，最后，通过决策分析的过程指导生成待合成情感语音参数，并得到目标说话人待合成情感语音。如图 4.2 所示，为本系统框图。

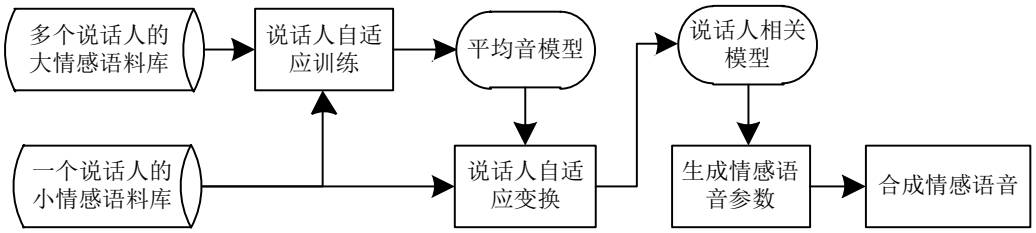


图 4.2 多情感说话人自适应训练系统框图

4.2.1 平均音模型

在基于隐 Markov 模型统计参数的情感语音合成系统中，为了训练得到优质的情感语音模型库，对情感语音数据的需求和要求都是比较严格的，如果采用单个说话人实现情感语音数据的录制，不仅需要耗费大量的时间和精力，情感语音数据的质量也得不到保证，可行性较低。但是，如果我们采用多个说话人共同进行情感语音数据库的搭建，不仅可以提高可行性，数据库的情感内容也更丰富。所以，本文选取了多个情感说话人进行情感语料库的搭建。

在基于隐 Markov 模型统计参数情感语音合成的训练过程中，首先，选取多个情感说话人的情感语音数据，分别对其基元进行训练，然后对所有基元的隐 Markov 模型进行概率分布统计，从而得到所有说话人情感语音数据的平均分布模型，即平均音模型[53]。

4.2.2 说话人自适应训练算法

为了提高合成情感语音质量，本文采用多个情感说话人语音数据训练得到平均音模型，由于情感说话人性别、性格、情感表达等方面的差异，声学模型会有较大偏差。为了避免因为说话人变化对训练模型所造成的影响，本文采用说话人自适应训练(Speaker Adaption Training, SAT)[51,53,54]的方法，对说话人差异进行归一化，以此提高模型的准确度，进而提高合成情感语音质量。如图 4.3 所示，为基于多个情感说话人语音数据的自适应训练的情感语音合成系统流程图。

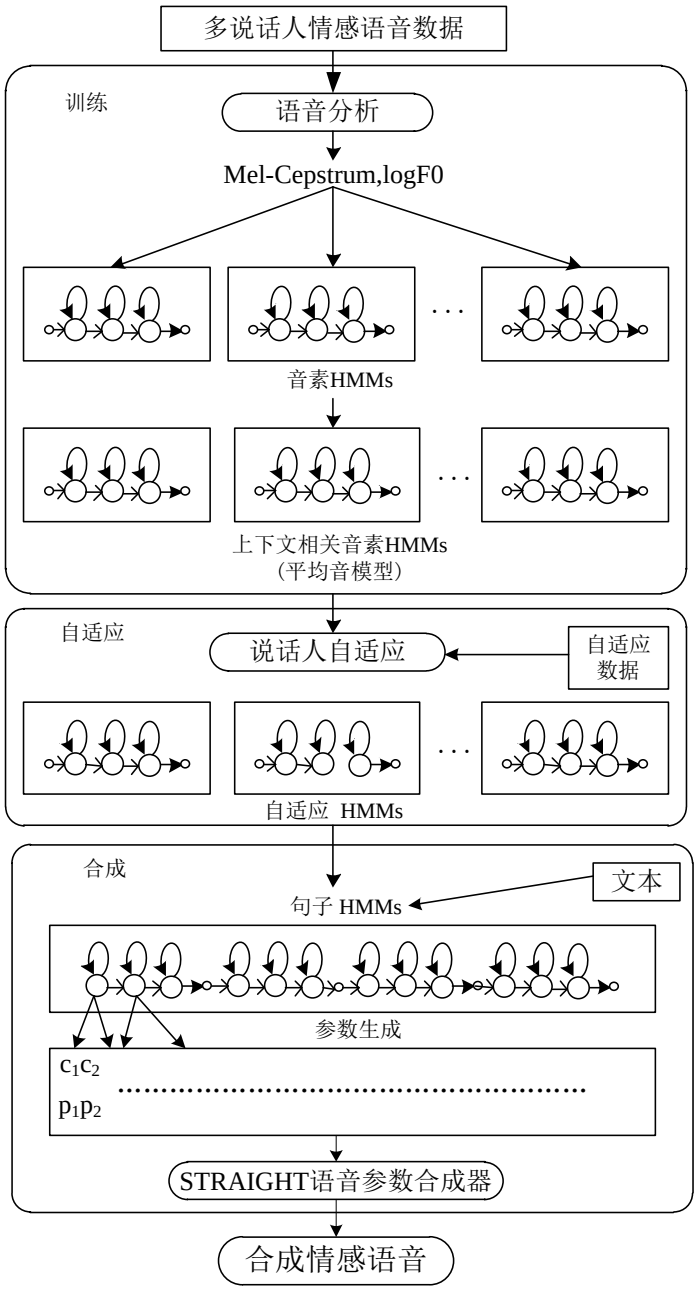


图 4.3 基于多情感说话人语音数据的自适应训练的情感语音合成流程图

由于在隐 Markov 模型统计参数语音合成系统中，我们需要采用多空间概率分布隐 Markov 模型(MSD-HMM)对基频部分进行建模。基于上述的上下文相关的 MSD-HSMM 语音合成单元，我们可以采用约束最大似然线性回归算法(constrained maximum likelihood linear regression, CMLLR)对多说话人情感语料库进行说话人自适应训练，从而获得多说话人情感语音的平均音模型。

首先，给定训练情感语音数据和目标说话人情感语音数据，为了反映两个模型之间差异，本文采用最大似然准则去估计两个模型数据之间的线性变换，并得到调整模型分布的协方差矩阵。在自适应训练过程中，我们需要对基频参数、频谱参数、时长参数等声学参数进行表征，并对这些参数的状态输出分布和时长分布进行估计、建模，但是最初的隐 Markov 模型没有对时长分布的精确描述，所以本文采用具有精确时长分布的半隐 Markov 模型(hidden semi-Markov model, HSMM)[51]对状态输出和时长分布进行同时控制建模，我们可以用一组线性回归方程，如公式（4.1）、公式（4.2）所示，来对说话人语音模型差异进行归一化处理：

$$\hat{o}_i^{(S)} = A^{(S)} o_i + b^{(S)} = W^{(S)} \xi_{(i)} \tag{4.1}$$

$$\hat{d}_i^{(s)} = \alpha^{(s)} d_i + \beta^{(s)} = X^{(s)} \xi_{(i)} \tag{4.2}$$

其中，公式（4.1）所示为状态输出分布变换方程，$\hat{o}_i$ 表示训练语音数据模型 s 的状态输出的均值向量，$W = [A, b]$ 为训练语音数据模型 s 的状态输出分布与平均音模型之间差异的变换矩阵，o_i 为其平均观测向量；公式（4.2）所示为状态时长分布变换方程，$\hat{d}_i$ 表示训练语音数据模型 s 的状态时长的均值向量。$X = [\alpha, \beta]$ 为训练语音数据模型 s 的状态时长分布与平均音模型之间差异的变换矩阵，d_i 为其平均时长，其中，$\xi = [o^T, 1]$。

然后，在进行完说话人自适应训练后，我们就可以利用待合成目标说话人的少量情感语句，采用 CMLLR 自适应算法对平均音模型进行说话人自适应变换，从而获得代表目标说话人的说话人自适应模型。在说话人自适应变换中，主要是利用说话人的状态输出和时长的概率分布的均值以及协方差矩阵，将混合语言平均音模型中的基频、频谱和时长参数变换为待合成语音的特征参数。如公式（4.3）所示为状态 i 下，特征向量 o 的变换方程，如公式（4.4）所示为状态 i 下，状态时长 d 的变换方程：

$$\begin{aligned} b_i(o) &= N(o; Au_i - b, A\Sigma_i A^T) \\ &= \left|A^{-1}\right| N(W\xi; u_i, \Sigma_i) \end{aligned} \tag{4.3}$$

$$\begin{aligned} p_i(d) &= N(d;\alpha m_i-\beta,\alpha\sigma_i^2\alpha) \\ &= \left|\alpha^{-1}\right| N(\alpha\psi;m_i,\sigma_i^2) \end{aligned} \tag{4.4}$$

其中，$\xi=[o^T,1]$，$\psi=[d,1]^T$，μ_i 为状态输出分布的均值，m_i 为时长分布的均值，Σ_i 为对角协方差矩阵，σ_i^2 为方差。$W=[A^{-1}\ \ b^{-1}]$ 为目标说话人状态输出概率密度分布的线性变换矩阵，$X=[\alpha^{-1},\beta^{-1}]$ 为状态时长概率密度分布的变换矩阵。

通过基于 HSMM 的自适应变换算法，可对语音声学特征参数进行归一化和特征处理。对于长度为 T 的自适应数据 $\boldsymbol{O}$，可对变换 $\Lambda=(\boldsymbol{W},\boldsymbol{X})$ 进行最大似然估计

$$\tilde{\Lambda}=(\tilde{W},\tilde{X})=\arg\max_{\Lambda} P(O\mid\lambda,\Lambda) \tag{4.5}$$

其中，λ 为 HSMM 的参数集。

当目标说话人数据量有限，不能满足每个模型分布都可以对应一个转换矩阵进行估计，这就需要多个分布共享一个转换矩阵，也就是回归矩阵的绑定[55]，最终可以通过采用较少的数据实现较好的自适应效果。如图 4.4 所示。

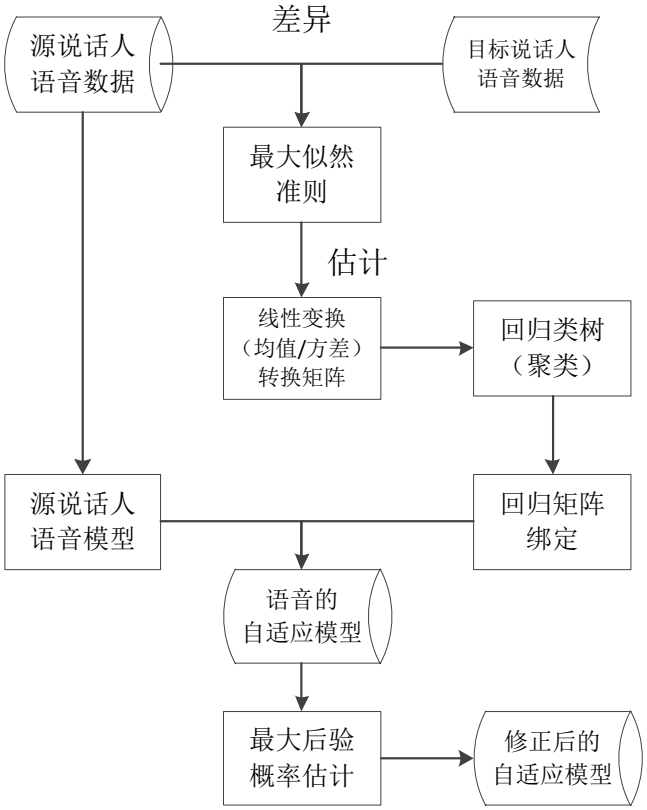


图 4.4 说话人自适应算法流程框图

最后，本文采用最大后验概率(Maximum A Posteriori, MAP)算法对模型进行修正和更新。对于给定的 HSMM 参数集 λ，假设其前向概率为 $\alpha_t(i)$，后向概率为 $\beta_t(i)$，在状态 i 下，其连续观测序列 $o_{t-d+1}\cdots\cdots o_t$ 的生成概率 $\kappa_t^d(i)$ 是：

$$\kappa_t^d(i)=\frac{1}{P(O|\lambda)}\sum_{\substack{j=1\\ j\neq i}}^{N}\alpha_{t-d}(j)p(d)\prod_{s=t-d+1}^{t}b_i(o_s)\beta_t(i) \tag{4.6}$$

最大后验概率估计描述如下：

$$\hat{u}_i=\frac{\omega\bar{u}_i+\sum_{t=1}^{T}\sum_{d=1}^{t}\kappa_t^d(i)\sum_{s=t-d+1}^{t}o_s}{\omega+\sum_{t=1}^{T}\sum_{d=1}^{t}\kappa_t^d(i)d} \tag{4.7}$$

$$\hat{m}_i=\frac{\tau\bar{m}_i+\sum_{t=1}^{T}\sum_{d=1}^{t}\kappa_t^d(i)d}{\tau+\sum_{t=1}^{T}\sum_{d=1}^{t}\kappa_t^d(i)d} \tag{4.8}$$

式中，$\bar{u}_i$ 和 $\bar{m}_i$ 代表线性回归变换之后的均值向量，ω 代表状态输出的 MAP 估计参数，而 τ 代表其时长分布 MAP 估计参数。$\hat{u}_i$ 和 $\hat{m}_i$ 代表自适应均值向量 $\bar{u}_i$ 以及 $\bar{m}_i$ 的加权平均 MAP 估计值。

4.2.3 基于说话人自适应训练的情感语音合成系统

本文提出了一种利用多情感说话人语音数据进行自适应训练实现情感语音合成的方法。利用多个情感说话人同一情感的语音数据作为训练语料，经过说话人自适应训练过程获得多说话人训练语音平均音模型；然后输入目标说话人的情感语音数据，通过说话人自适应转换过程[56]获得目标说话人该情感的相关模型，从而合成得到情感语音，其方法流程图如图 4.5 所示。

与传统的基于隐 Markov 模型的语音合成方法相比，本论文在训练阶段加入了说话人自适应训练过程，获得多个说话人的情感语音平均音模型，通过此方法，可以减小语音库中说话人的差异所造成的影响，提高合成语音的情感相似度，在平均音模型的基础上，通过说话人自适应变换算法，只用少量的待合成的情感语料，就能够合成出自然度、流利度、情感相似度都很好的情感语音。

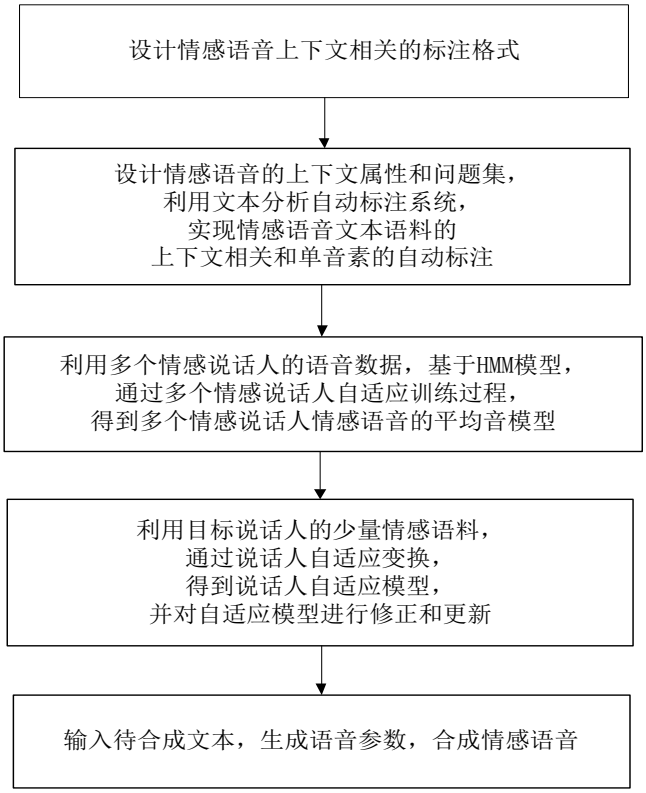


图 4.5 情感语音合成系统的方法流程图

如图 4.6 所示为基于多个情感说话人语音数据自适应训练方法，实现情感语音合成的系统框图，本系统可分为训练阶段、自适应阶段、合成阶段三部分。

训练阶段，给定多说话人目标情感的大型情感语音数据库和目标说话人目标情感的小型语音数据库，其中语音数据文件经过 STRAIGHT 参数提取过程，得到训练模型所需的基频、谱参数等声学参数文件，文本文件经过文本分析过程，由标注生成程序得到包含音素信息的单音素标注文件和包含上下文语境信息的上下文相关的标注文件。然后在上下文属性和问题集的指导下，对基元模型进行 HMM 训练，并通过决策树聚类得到 HMM 模型库。

自适应阶段，首先采用约束最大似然线性回归算法对多说话人情感语音数据模型进行说话人自适应训练，从而获得多说话人情感语音数据的平均音模型。之后，在目标说话人目标情感语音数据的指导下，同样采用约束最大似然线性回归算法对平均音模型进行说话人自适应变换，得到说话人相关的自适应模型，最后采用最大后验概率对自适应模型进行修正和更新。

合成阶段，和基于 HMM 的统计参数语音合成方法原理相同，首先输入待合成目标说话人目标情感的语音文本，然后通过文本分析过程，由标注生成程序生成得到目标文本的上下文相关的标注文件。在自适应模型的指导下，通过决策分析得到

目标语音的上下文相关的 HMM 决策序列，并生成相应的语音参数。最后采用 STRAIGHT 语音合成器，合成得到目标说话人目标情感的语音。

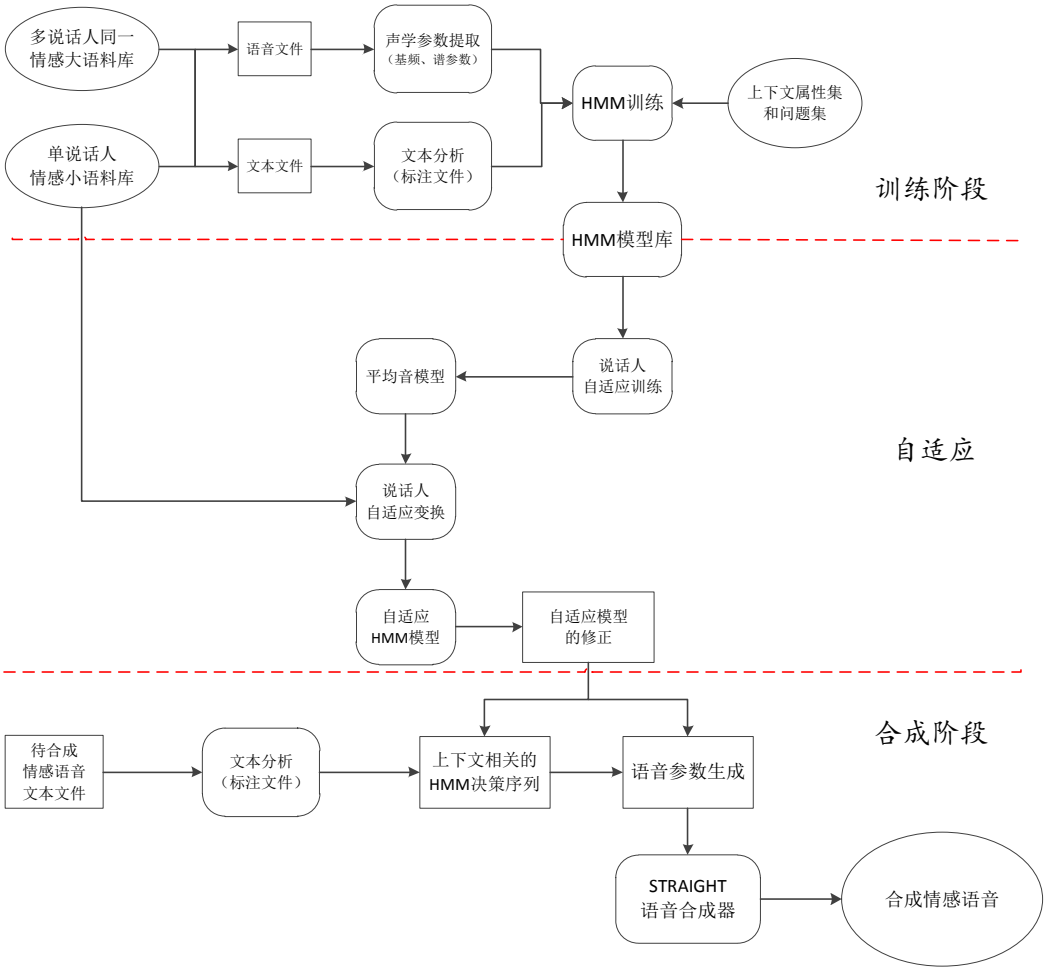


图 4.6 基于多情感说话人语音数据自适应训练的情感语音合成系统

4.3 本章小结

首先，本章分析了基于隐 Markov 模型的参数语音合成方法，构建了说话人相关的统计参数语音合成系统，并对基于隐 Markov 模型的统计参数语音合成方法及合成效果进行特点分析；之后，本章重点介绍了通过采用多个情感说话人语音数据进行自适应训练的方法，实现情感语音合成的相关原理，并搭建了基于多情感说话人语音数据自适应训练的情感语音合成系统，本系统只需目标说话人少量的语音数据，就可以合成出自然度和情感相似度都比较高的情感语音。

第 5 章 实验及测评

本章主要包含两个内容的实验及评测。实验一，针对汉语统计参数语音合成中的上下文相关标注生成，设计了以声韵母为合成基元包含六层的上下文相关的标注格式。搭建了基于隐 Markov 模型的统计参数语音合成系统，并通过主、客观实验评测不同标注信息对合成语音音质的影响。实验二，分别采用单情感说话人自适应训练的统计参数语音合成方法和多情感说话人语音数据自适应训练的统计参数语音合成方法得到情感语音，并通过 MOS 和 EMOS 评测方法对得到的情感语音进行自然度和情感相似度分析。

5.1 实验评测方法

本文的实验测评采用主观测评和客观测评[57]两种测评方法，对合成语音效果进行综合评估。

5.1.1 客观评测

通过将合成语音参数和原始语音文件参数进行对比，计算相应的误差进行分析。本文主要根据公式（5.1）对语音文件基元（音素）时长、语句时长、基频、谱质心等参数进行了均方根误差分析。

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} W_i^2} \tag{5.1}$$

式中，N 为对比语音文件个数，W_i 为合成语音和原始语音的参数误差。其中，基元（音素）时长的误差W 定义如下：

$$W = \left|(T_2 - T_1) - (t_2 - t_1)\right| \tag{5.2}$$

式中，t_1 为标注文件下得到的合成语音基元的起始时间，t_2 为标注文件下得到的合成语音基元的截止时间；T_1 为原始语音基元的起始时间，T_2 为原始语音基元的截止时间。

语句时长的误差W 定义如下：

$$W = \left|(T_e - T_0) - (t_e - t_0)\right| \tag{5.3}$$

式中，t_0 为合成语句的起始时间，t_e 为合成语句的截止时间；T_0 为原始语句的起始时间，T_e 为原始语句的截止时间。

基频是语音韵律的重要体现，是分析合成语音韵律特征的一个重要参数。基频的误差W 定义如下：

$$W = \left| f_2 - f_1 \right| \tag{5.4}$$

式中，f_1为合成语句的基频均值；f_2为原始语句的基频均值。

谱质心(Spectral Centroid)[58]是在一定频率范围内通过能量加权平均的频率，其计算公式如下：

$$SC = \frac{\sum_{n=1}^{N} f(n) \bullet E(n)}{\sum_{n=1}^{N} E(n)} \tag{5.5}$$

式中，N 为 DFT 的长度，$f(n)$为离散信号经短时傅里叶变化后对应的频率，$E(n)$为离散信号经短时傅里叶变化后对应频率的谱能量，SC 为信号的谱质心。

谱质心的误差W定义如下：

$$W = \left| SC_2 - SC_1 \right| \tag{5.6}$$

式中，SC_1为合成语句的谱质心；SC_2为原始语句的谱质心。

5.1.2 主观评测

为了进一步对合成语音的质量进行评估，本文采用了平均意见得分(Mean Opinion Score, MOS)的主观测评方法，对合成语音的自然度进行评估。通过给 22 名普通话评测者随机播放 250 句（10 句*25 组）合成的语音，评测过程中，评测者根据播放语音的先后顺序进行打分，评测者根据合成语音的自然度，对每句合成语音按 5 分制进行打分，打分标准如表 5.1 所示。

表 5.1 MOS 评测分值标准表

分值	评测标准
0-1	劣，极差，听不懂
1-2	差，勉强，听不大清
2-3	中，有延迟，可以接受
3-4	良，听得清楚，愿意接受
4-5	优，很自然

为了进一步对合成情感语音的相似度进行评估，本文又采用了情感相似度平均意见得分(Emotional Mean Opinion Score, EMOS)的主观测评方法，对合成语音的情感相似度进行评估。通过给 22 名情感语音评测者随机播放 250 句（10 句*25 组）合成的语音，评测过程中，评测者根据播放语音的先后顺序进行打分，评测者根据合成情感语音的情感度，对每句合成语音按 5 分制进行打分，打分标准如表 5.2 所示。

表 5.2 EMOS 评测分值标准表

分值	评测标准
0-1	劣，情感度不明
1-2	差，情感度模糊
2-3	中，情感度可以接受
3-4	良，情感度愿意接受
4-5	优，情感相似度理想

5.2 实验结果分析

5.2.1 上下文相关的标注格式设计实验

图 5.1 所示为上下文相关的标注格式设计实验框图，本实验针对携带不同语境信息的上下文相关的标注文件，采用基于 HMM 的统计参数语音合成系统，如图 4.1 所示，并对不同标注格式下得到的合成语音进行主客观评测对比分析，验证不同层次标注格式信息对合成语音音质的影响，最终确定面向 HMM 统计参数语音合成系统的上下文相关的标注格式，并通过情感语音数据对该标注格式在情感语音合成实验可行性进行实验验证。

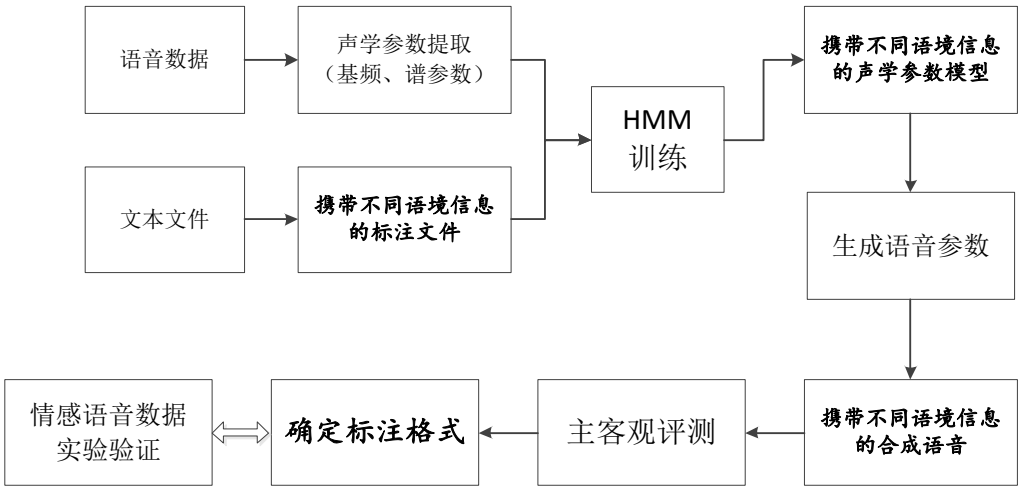


图 5.1 标注格式生成设计实验方案框图

图 5.2 所示为上下文相关的标注格式设计实验方案详细流程图，本实验方案共包括训练阶段、合成阶段、评测阶段三部分内容。

训练阶段，输入普通话语音文件，采用 STRAIGHT 参数提取算法，提取语音文件的基频、非周期索引、Mel 倒谱等声学参数信息，同时，输入对应的普通话文本文件，通过文本分析过程，得到其单音素的标注文件，并采用不同标注格式，分别

得到携带不同语境信息的上下文相关的标注文件，之后，在上下文属性和问题集的指导下，进行 HMM 训练，并做决策树聚类，生成 HMM 模型库。

合成阶段，输入待合成的普通话语音文本文件，通过文本分析过程，得到携带不同语境信息的上下文相关的标注文件，之后，在 HMM 模型库的指导下，进行决策分析，生成上下文相关的 HMM 决策序列，并通过参数生成算法得到待合成语音声学参数，最后，通过 STRAIGHT 参数合成器，得到携带不同语境信息的上下文相关的标注文件指导下，合成的普通话语音。

评测阶段，分别采用主观评测和客观评测两种方法，对不同标注格式下得到的合成语音进行对比分析，最终确定面向普通话统计参数语音合成的标注格式。其中，主观评测采用 MOS 评测方法，对合成语音的自然度进行对比分析；客观评测采用 RMSE 评测方法，对合成语音和原始语音的基频、时长、谱质心等声学参数进行均方根误差对比分析。之后，采用情感语音训练数据，对基于统计参数语音合成方法的情感语音合成实验进行可行性验证。

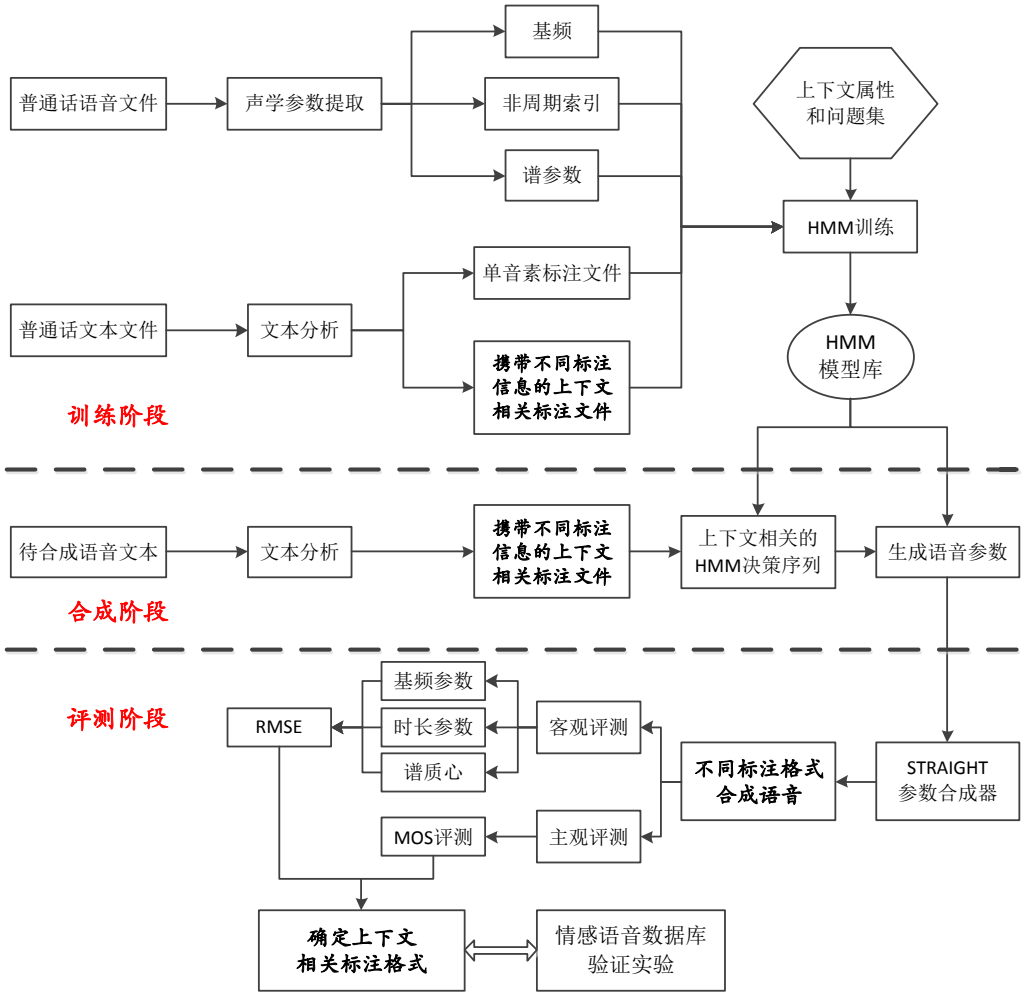
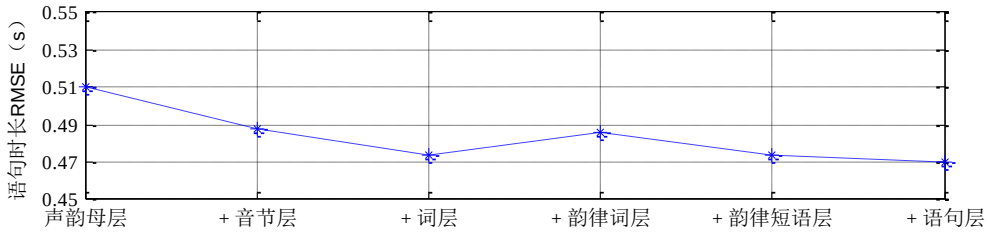
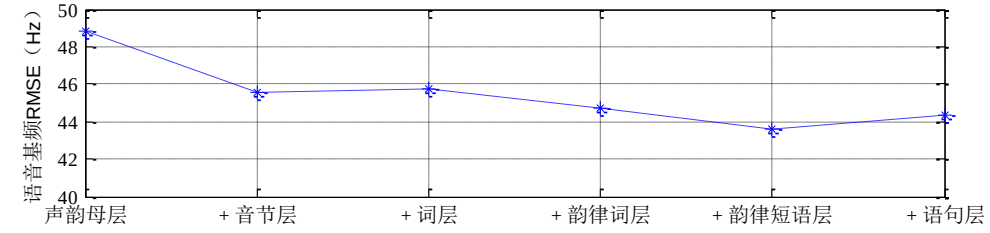


图 5.2 标注格式生成设计对比实验方案详细流程图

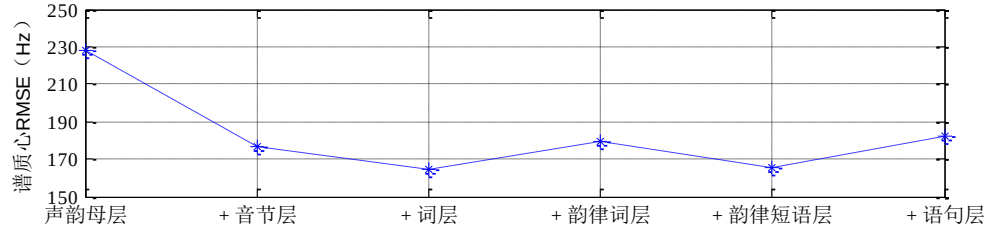
如图 5.3 所示为随着增加不同层次的标注信息后，得到的合成语音参数的 RMSE 对比结果。图(a)中显示了包含不同标注信息的情况下，语音时长的 RMSE 对比情况，如图所示，随着标注信息的逐渐增加，时长的误差是逐渐减少的，虽然在增加韵律词层信息后，合成语音时长误差略微增大，但在增加了韵律短语层和语句层信息后，时长误差得到有效改善；图(b)中显示了语音基频的 RMSE 对比情况，随着标注信息的逐渐增加，除语句层标注信息对合成语音造成较小误差外，基频误差总体呈下降趋势；图(c)中显示了语音谱质心的 RMSE 对比情况，可以看出，随着标注信息量增加，谱质心误差明显下降，只是在增加韵律词层信息和语句层信息后，合成语音谱质心误差会略微增加。从客观评测结果可以看出，随着标注层的增加，合成语音参数误差总体呈减小的趋势。因此，本文设计的标注文件符合汉语统计参数语音合成系统的标注要求，对合成语音的客观参数要求起到良好的效果。



（a）依次增加各层信息后语句时长RMSE对比图例



（b）依次增加各层信息后语音基频RMSE对比图例



（c）依次增加各层信息后谱质心RMSE对比图例

图 5.3 客观评测结果

如图 5.4 所示为随着增加不同层次的标注信息后，得到的合成语音的主观评测结果。很明显，在增加音节层信息后，合成语音的 MOS 评分明显提高；随着增加各标注层的标注信息，得到的合成语音自然度得分逐渐升高；最终，在完整的标注

信息下，得到的合成语音自然度最好。通过主观评测结果可以看出，在包含六层标注信息下得到的合成语音自然度最好。

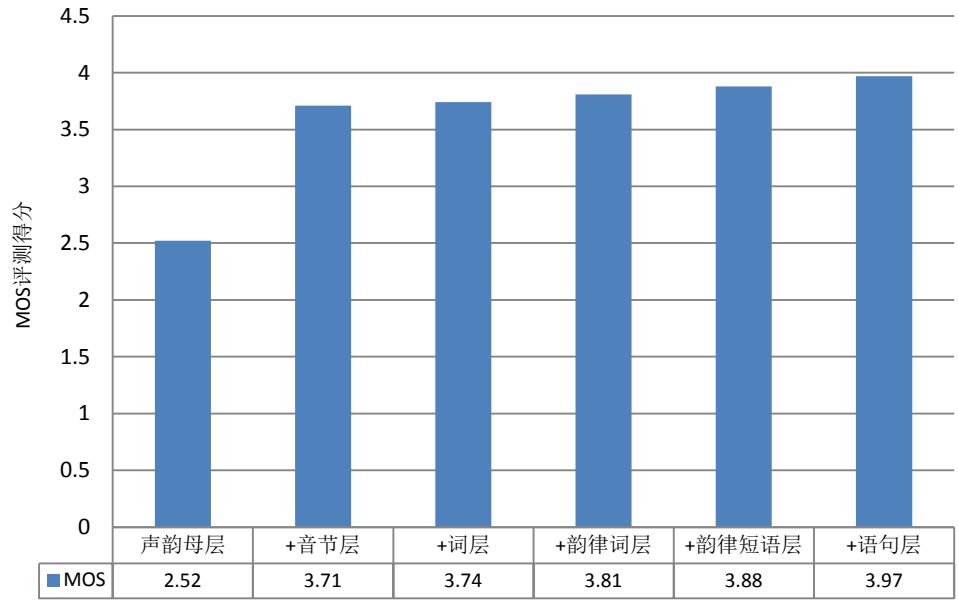


图 5.4 主观评测结果

实验结果表明，随着标注信息的逐渐完善，合成语音参数误差总体呈减小的趋势，合成语音自然度得分逐渐升高，最终，在完整的标注信息下，得到的合成语音自然度最好，因此，本文设计的标注文件符合汉语统计参数语音合成系统的标注要求，对合成语音的客观参数要求起到良好的效果。同时，本文设计的上下文相关的6层标注格式满足情感语音合成实验需求。

5.2.2 情感语音合成实验

图 5.5 所示为情感语音合成实验方案系统框图，为了评估合成的情感语音，本文训练了 2 种不同的 MSD-HSMM 模型。其中，SAT2 为本文提出的情感语音合成实验方案模型，SAT1[8]为对比实验方案模型：

1) SAT1 模型：本文首先采用 5000 句的中性语音数据库作为训练集，通过说话人自适应训练获得平均音模型，然后对分别输入目标说话人目标情感语音数据作为自适应语句，并逐个进行实验，对平均音模型进行说话人自适应变换，并获得目标情感说话人目标情感的自适应模型。

2) SAT2 模型：实验方法和 SAT1 模型搭建过程相同，对平均音模型进行说话人自适应变换，并获得目标说话人目标情感的自适应模型。不同的是，SAT2 采用和 SAT1 不同的训练语音数据库（多个情感语音说话人的情感语音训练集）。

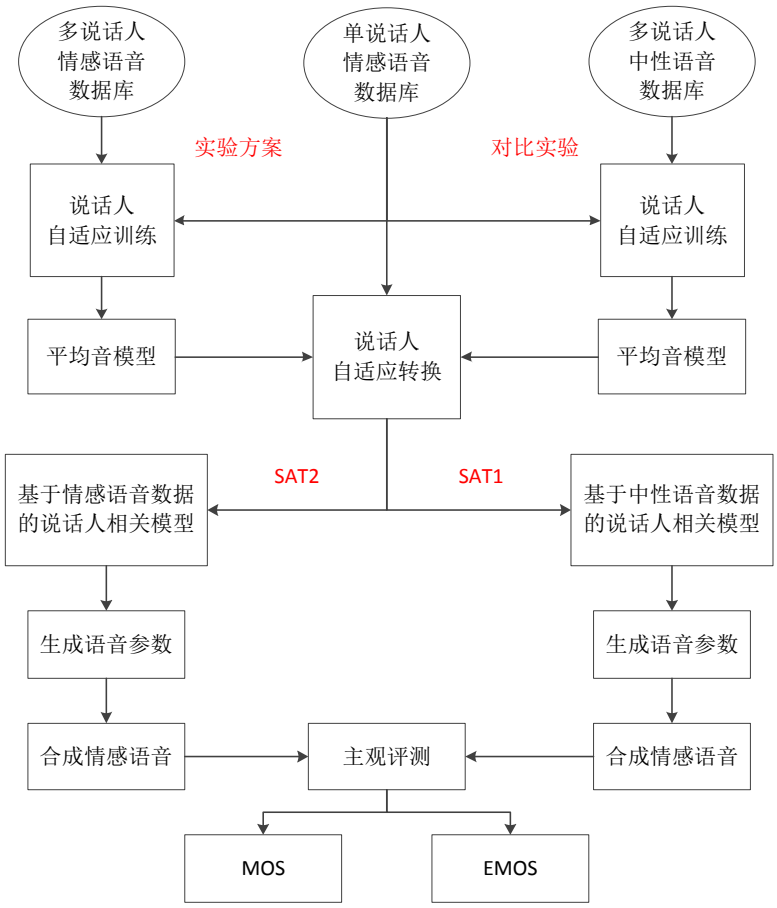


图 5.5 情感语音合成实验方案框图

然后，输入待合成情感语音文本，在决策树问题集的指导下进行决策分析，从相应的训练语音模型库中挑选合适的基元模型，通过参数生成器合成得到目标情感语音， SAT1 和 SAT2 模型分别合成出每种情感的 10 句情感测试语句，共计 220 句情感合成语句。最后，本文分别采用 MOS 评测和 EMOS 评测的主观评测方法，对基于两种训练数据的语音模型下，合成得到的情感语音的自然度和情感相似度进行对比分析。如图 5.6 所示，为本实验方案的详细流程图。

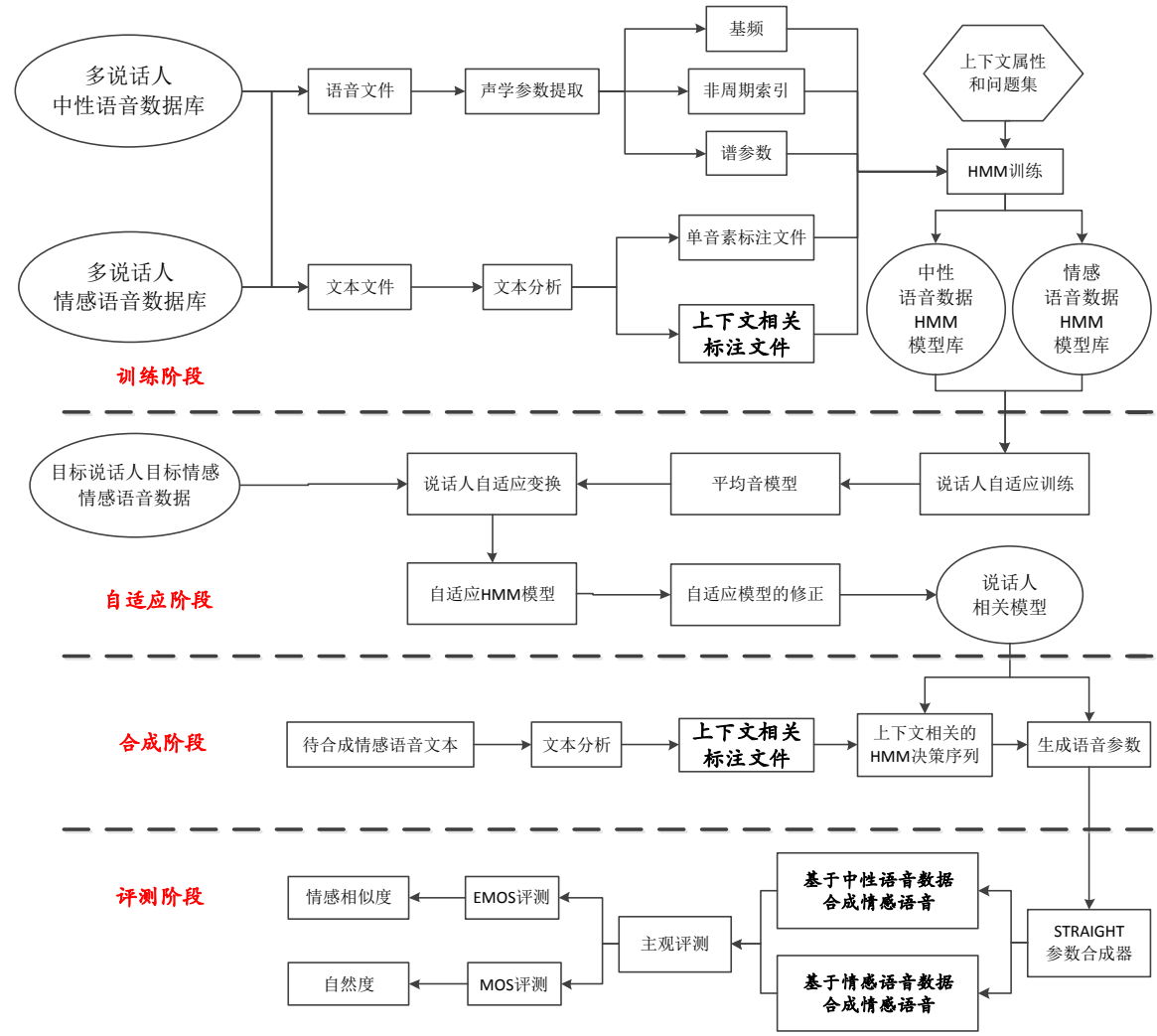


图 5.6 情感语音合成实验方案详细流程图

如图 5.7 所示为两种训练模型下得到的合成情感语音的 MOS 得分。其中，采用 SAT1 模型得到的情感语音的 MOS 得分最低为 2.1，最高为 3.4，平均值为 2.4；采用 SAT2 模型得到的情感语音的 MOS 得分最低为 3.0，最高为 4.1，平均值为 3.5。可以看出，采用 SAT2 模型得到合成情感语音自然度明显优于采用 SAT1 模型合成得到的情感语音自然度。

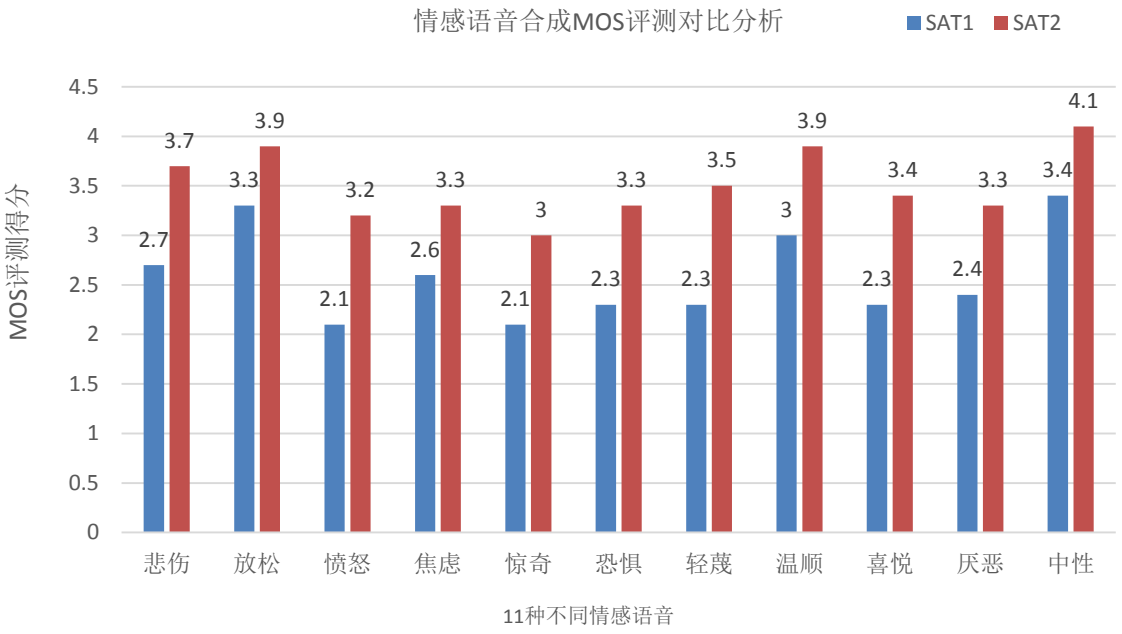


图 5.7 情感语音合成 MOS 评测对比分

同时，本文还采用 EMOS 评测方法对两种训练模型得到合成语音的情感相似度进行对比分析。如图 5.8 所示为两种训练模型下得到的合成情感语音的 EMOS 评测得分。其中，采用 SAT1 模型得到的情感语音的 EMOS 得分最低为 2.0，最高为 3.5，平均值为 2.6；采用 SAT2 模型得到的情感语音的 EMOS 得分最低为 3.0，最高为 4.0，平均值为 3.4。可以看出，采用 SAT2 模型得到合成语音的情感相似度明显优于采用 SAT1 模型合成语音的情感相似度。

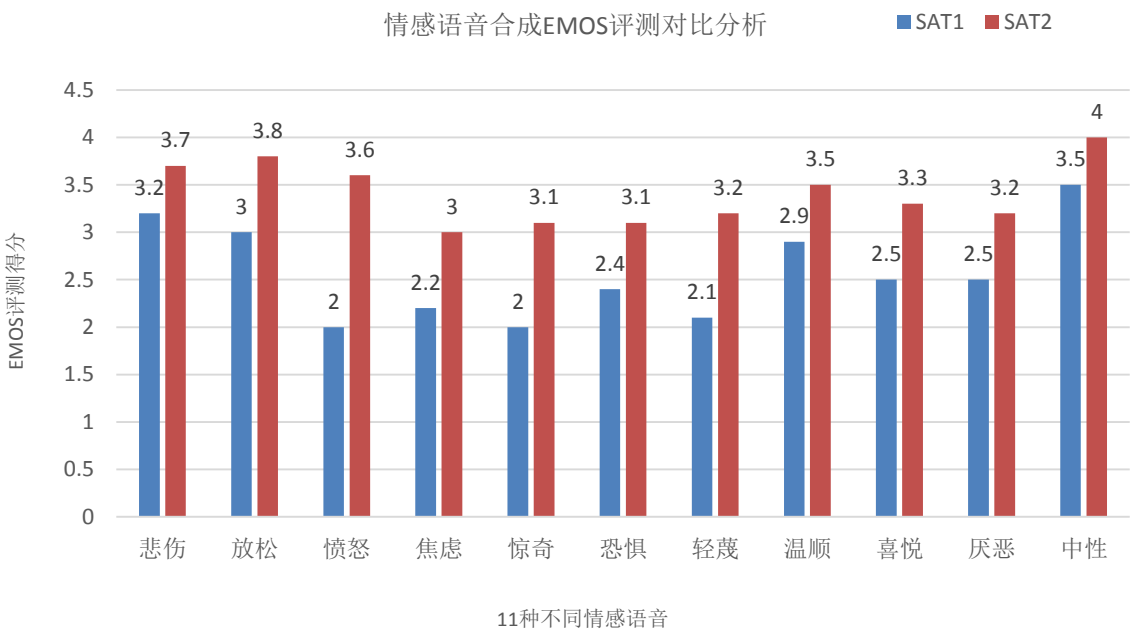


图 5.8 情感语音合成 EMOS 评测对比分析

实验结果表明，采用基于多情感说话人语音数据的自适应方法合成得到的情感语音，自然度和情感相似度得分都有明显提高。

5.3 本章小结

本章首先分别介绍了主观评测、客观评测两种语音质量评估方法，并分别通过主、客观评测验证了不同标注信息对合成语音音质的影响，实验结果表明，随着标注信息的逐渐完善，合成语音参数误差总体呈减小的趋势，合成语音自然度得分逐渐升高，最终，在完整的标注信息下，对合成语音的客观参数要求起到良好的效果，得到的合成语音自然度最好，因此，本文设计的标注文件符合汉语统计参数语音合成系统的标注要求；之后，本章分别采用基于单情感说话人语音数据进行自适应训练的统计参数语音合成方法和基于多情感说话人语音数据进行自适应训练的统计参数语音合成方法得到情感语音，并通过 MOS 和 EMOS 评测方法对得到的情感语音进行自然度和情感相似度分析，实验结果表明，采用基于多情感说话人语音数据的自适应方法合成得到的情感语音，自然度和情感相似度得分都有明显提高。

第 6 章 总结及展望

目前，随着语音合成技术的研究与发展，合成语音音质得到较大提升，但当前语音合成技术的研究仍以中性化语音为主，对情感语音合成的研究较少。将来，人类生活对智能语音的需求不仅要涵盖基本的文字内容，还要承载丰富的情感信息，情感语音合成的研究将是智能语音研究领域的必然趋势。

6.1 论文总结

本文建立了一个多说话人多种情感的语音数据库，并针对汉语统计参数语音合成中的上下文相关标注生成，设计了包含 6 层上下文相关的标注格式，在此基础上，采用多说话人的情感语音数据，基于说话人自适应训练的统计参数语音合成方法，实现了情感语音合成。论文的主要工作和创新如下：

1.建立了多个说话人多种情感的情感语料库。本工作在高隔音的情感语音专业录影棚中进行，采用诱发方式对录音人进行情感获取，本论文共采集了 7 个男性说话人和 7 个女性说话人的普通话情感语音数据，其中每个说话人的语音数据包含 11 种典型情感，录制得到的情感语音数据以 Microsoft WAV 格式（单通道、16bit、16kHz 采样频率）进行保存。

2.实现了面向统计参数语音合成的标注生成。本文针对汉语统计参数语音合成中的上下文相关标注生成，设计了一套包含 6 层上下文相关信息的标注格式，并采用以声韵母为合成基元的基于隐 Markov 模型(Hidden Markov Model, HMM)的统计参数语音合成系统，通过主、客观实验评测了不同标注信息对合成语音音质的影响。实验结果表明，本文设计的上下文相关的六层标注格式满足情感语音合成实验需求。

3.提出了一种基于多个情感说话人的说话人自适应训练(Speaker Adaptive Training, SAT)的情感语音统计参数合成方法。首先通过说话人自适应训练得到多个说话人情感语音的平均音模型，再给定目标说话人的情感语音数据，经过说话人自适应变换，得到目标说话人目标情感的语音声学模型，并通过参数生成过程，最后合成得到目标说话人的情感语音。实验结果表明，本方法合成得到的情感语音具有较高的自然度和情感相似度。

6.2 下一步工作展望

本文针对汉语的统计参数语音合成方法设计了上下文相关的标注格式，并扩建了情感语料库，通过采用基于多个情感说话人自适应训练的方法，实现情感语音合成。合成语音虽然在自然度和情感相似度方面都得到明显改善，但通过系统实验证

明，情感语音合成的工作还有很多问题有待改进。下一步工作主要可从以下几个方面展开：

1.连续型情感语音文本标注格式设计。相比离散的情感语音文本标注方法，连续型的情感语音文本标注方法更能体现人类情感的平滑变化的特性，如何设计一套合理的情感语音文本标注格式，并与上下文相关的文本标注格式，共同通过文本分析过程，得到语境信息标注文件，将作为下一步工作进行研究。

2.继续扩建情感语音数据库。数据库的大小和语音数据质量，对统计参数语音合成系统合成语音效果会产生重要影响，人类的情感复杂多变，11 种情感语音远远不能代表人类的情绪表达，如何创建一个情感色彩更加丰富、语音数据更加充实的情感语音数据库，将是一个长久而又必要的研究课题。

3.采用深度学习的方法进行情感语音合成的研究。随着深度神经网络在智能人机交互领域的广泛应用，基于深度学习方法的情感语音合成成为语音信号处理领域的重要研究内容，实验室下一步将会采用递归神经网络(Recurrent Neural Network, RNN)方法实现情感语音合成，以期有所突破。

随着智能人机交互技术的迅速发展，语音信号处理领域的相关研究成果逐渐深入到人们的生活、工作和学习中去，情感语音合成的研究必将成为当下科研单位的研究热点之一，其研究成果必然会给人类生活带来更大的改善。

参考文献

[1] 戴礼荣, 张仕良. 深度语音信号与信息处理: 研究进展与展望[J]. 数据采集与处理, 2014, 29(2): 171-179.

[2] 王坚, 张媛媛. 基于深度神经网络的汉语语音合成的研究[J]. 计算机科学, 2015, 1.

[3] Dutoit T. An introduction to text-to-speech synthesis[M]. Springer, 1997.

[4] Yamagishi J, Tamura M, Masuko T, et al. A training method of average voice model for HMM-based speech synthesis[J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2003, 86(8): 1956-1963.

[5] 王海燕, 杨鸿武, 甘振业, 等. 基于说话人自适应训练的汉藏双语语音合成[J]. 清华大学学报:自然科学版, 2013(6):776-780.

[6] Yang H, Oura K, Wang H, et al. Using speaker adaptive training to realize Mandarin-Tibetan cross-lingual speech synthesis[J]. Multimedia Tools and Applications, 2015, 74(22):9927-9942.

[7] 王晓丽. 高表现力语音声学建模的研究[D]. 兰州: 西北师范大学, 2011.

[8] 鲁小勇. 情感语音合成的研究[D]. 兰州: 西北师范大学, 2013.

[9] Moulines E, Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones[J]. Speech communication, 1990, 9(5-6): 453-467.

[10] Valbret H, Moulines E, Tubach J P. Voice transformation using PSOLA technique[J]. Speech Communication, 1992. 11: 175-187.

[11] Iida A, Campbell N, Higuchi F, et al. A corpus-based speech synthesis system with emotion[J]. Speech Communication, 2003, 40(1): 161-187.

[12] 刘艳. 普通话的情感语音韵律分析[D]. 南京: 南京师范大学, 2011.

[13] Zen H, Tokuda K, Black A W. Statistical parametric speech synthesis[J]. Speech Communication, 2009, 51(11):1039-1064.

[14] Tokuda K, Yoshimura T, Masuko T, et al. Speech parameter generation algorithms for HMM-based speech synthesis[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000:1315-1318.

[15] Masuko T. Multi-Space Probability Distribution HMM[J]. IEICE Transactions on Information and Systems, 2002, E85-D(3): 455-464.

[16] 雷鸣. 统计参数语音合成中的声学模型建模方法研究[D]. 合肥: 中国科学技术大学, 2012.

[17] Kawahara H, Masuda-Katsuse I, Cheveigné A D. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds 1[J]. Speech Communication, 1999, 27(3-4):187-207.

[18] 陈浩, 师雪姣, 肖智议, 等. 高表现力情感语料库的设计[J]. 计算机与数字工程, 2014, 42(8): 1383-1385.

[19] 徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析[J]. 中文信息学报, 2008, 22(1): 116-122.

[20] Gajšek R, Štruc V, Vesnicer B, et al. Analysis and Assessment of AvID: Multi-Modal Emotional Database.[C]// International Conference on Text, Speech and Dialogue. Springer-Verlag, 2009:266-273.

[21] McGilloway S, Cowie R, Douglas-Cowie E, et al. Approaching automatic recognition of emotion from voice: a rough benchmark[C]// ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, 2000.
[22] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech.[C]// European Conference on Speech Communication and Technology, Lisbon, Portugal, September, 2005:1517-1520.
[23] 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014, 25(1): 37-50.
[24] 刘星星. 基于曲线回归分析的情感语音合成[D]. 太原: 太原理工大学, 2014.
[25] 孟昭兰. 情绪心理学[M]. 北京:北京大学出版社,2005.
[26] 梁宁建. 心理学导论[M]. 上海:上海教育出版社,2006.
[27] Cowie R, Cornelius R R. Describing the emotional states that are expressed in speech[J]. Speech Communication, 2003, 40(s 1–2):5-32.
[28] 蒋丹宁, 蔡莲红. 基于韵律特征的汉语情感语音分类[C]// 中国情感计算及智能交互学术会议. 2003.
[29] Paul Ekman. An argument for basic emotions[J]. Cognition and Emotion, 1992, 6(3-4):169-200.
[30] 李虎孬. 情感语音合成之语料库的创建[D]. 乌鲁木齐: 新疆大学, 2014.
[31] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction[J]. Signal Processing Magazine, 2001, 18(1): 32-80.
[32] Plutchik R. The Multifactor-Analytic Theory of Emotion[J]. Journal of Psychology Interdisciplinary and Applied, 1960, 50(1):153-171.
[33] Wang R H, Liu Q, Tang D. A new Chinese text-to-speech system with high naturalness[C]// IEEE Fourth International Conference on Spoken Language Proceedings, 1996, 3:1441-1444.
[34] 袁里驰. 基于统计的句法分析方法[J]. 中南大学学报: 自然科学版, 2014, 45(8): 2669-2675.
[35] 丁蓉. 自动语义标注方法研究[D]. 兰州: 兰州理工大学, 2012.
[36] 杨舟. 基于自然语言处理的专利文档自动语义标注方法研究[D]. 杭州: 浙江大学, 2011.
[37] 杨鸿武, 王晓丽, 陈龙, 等. 基于语法树高度的汉语韵律短语预测[J]. 计算机工程与应用, 2010, 46(36):139-143.
[38] 杨鸿武, 朱玲. 基于句法特征的汉语韵律边界预测[J]. 西北师范大学学报: 自然科学版, 2013, 49(1): 41-45.
[39] 杨辰雨. 语音合成音库自动标注方法研究[D]. 合肥: 中国科学技术大学, 2014.
[40] 李华栋, 贾真, 尹红风, 等. 基于规则的汉语兼类词标注方法[J]. 计算机应用, 2014, 34(8): 2197-2201.
[41] 蔡莲红. 现代语音技术基础与应用[M]. 北京: 清华大学出版社, 2003.
[42] Chen Z, Hu G, Wang X. Text Normalization in Chinese Text-to-Speech System[J]. Chinese Information Processing, 2003, 17(4): 45-51.
[43] 谭同超. 有限状态机及其应用[D]. 广州: 华南理工大学, 2013.
[44] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 1996, 22(1): 39-71.
[45] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2008.
[46] 郑敏, 蔡莲红. 一种新的基于规则的多音字自动注音方法[J]. 第二届全国学生计算语言学研讨会论文集, 北京, 2004: 238-243.

[47] Rendel A, Sorin A, Hoory R, et al. Towards automatic phonetic segmentation for TTS[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing, 2012:4533-4536.

[48] 陈龙, 杨鸿武, 蔡莲红. 基于 TBL 算法的汉语韵律词预测[J]. 西北师范大学学报: 自然科学版, 2008, 44(1): 47-51.

[49] 梁青青, 杨鸿武, 郭威彤, 等. 利用五度字调模型实现普通话到兰州方言的转换[J]. 声学技术, 2010 (6): 620-625.

[50] Valentini-Botinhao C, Yamagishi J, King S. Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing, 2011:5112-5115.

[51] Yamagishi J, Kobayashi T. Average-Voice-Based Speech Synthesis Using HSMM-Based Speaker Adaptation and Adaptive Training[J]. IEICE Transactions on Information and Systems, 2007, E90D(2):533-543.

[52] 凌震华. 基于统计声学建模的语音合成技术研究[D]. 合肥: 中国科学技术大学, 2008.

[53] Yamagishi J. Average-Voice-Based Speech Synthesis[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006, 1.

[54] Yamagishi J, Kobayashi T, Nakano Y, et al. Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm[J]. IEEE Transactions on Audio Speech and Language Processing, 2009, 17(1):66-83.

[55] 赵欢欢. 基于隐马尔可夫模型的说话人转换研究[D]. 合肥: 中国科学技术大学, 2009.

[56] 宋文龙. 基于说话人自适应训练的统计参数语音合成的研究[D]. 兰州: 西北师范大学, 2013.

[57] 徐世鹏, 杨鸿武, 王海燕. 面向藏语语音合成的语音基元自动标注方法[J]. 计算机工程与应用, 2015 (6): 199-203.

[58] Park T H, Adviser-Lansky P, Adviser-Cook P. Towards automatic musical instrument timbre recognition[M]. Princeton University, 2004.

附录 A

附表 1 普通话词性表

词性	POS	符号表示
名词	noun	n
时间	time	t
地点	place	s
方位	direction	f
动词	verb	v
形容词	adjective	a
区分词	distinguish	b
状态词	state	z
代词	pronouns	r
数词	numeral	m
数量词	quantifier	q
副词	adverb	d
介词	preposition	p
连接词	conjunction	c
小品词	particle	u
感叹词	interjection	e
语气词	modal particles	y
象声词	onomatopoeia	o
前缀	prefix	h
后缀	suffix	k
字串	string	x
标点符号	punctuation	w

附录 B

附表 2 上下文相关标注字母含义表

标注层	字母	含义
声韵母层	p1	前前音素
	p2	前音素
	p3	当前音素
	p4	后音素
	p5	后后音素
	p6	当前音素在当前音节的位置（向前）
	p7	当前音素在当前音节的位置（向后）
音节层	a1	前一音节的首音素
	a2	前一音节的末音素
	a3	前一音节在词典中的声调类型
	a4	前一音节在文本分析中的声调类型
	a5	前一音节的音素个数
	b1	当前音节的首音素
	b2	当前音节的末音素
	b3	当前音节在词典中的声调类型
	b4	当前音节在文本分析中的声调类型
	b5	当前音节的音素个数
	b6	当前音节到在当前词的位置（向前）
	b7	当前音节到在当前词的位置（向后）
	b8	当前音节到在当前韵律词的位置（向前）
	b9	当前音节到在当前韵律词的位置（向后）
	b10	当前音节到在当前短语的位置（向前）
	b11	当前音节到在当前短语的位置（向后）
	c1	后一音节的首音素
	c2	后一音节的末音素
	c3	后一音节在词典中的声调类型
	c4	后一音节在文本分析中的声调类型
	c5	后一音节的音素个数

词层	d1	前一词的词性
	d2	前一词的音节个数
	e1	当前词的词性
	e2	当前词的音节个数
	e3	当前词到在当前韵律词的位置（向前）
	e4	当前词到在当前韵律词的位置（向后）
	f1	后一词的词性
	f2	后一词的音节个数
韵律词层	g1	前一韵律词的音节个数
	g2	前一韵律词的词个数
	h1	当前韵律词的音节个数
	h2	当前韵律词的词个数
	h3	当前韵律词到在当前短语的位置（向前）
	h4	当前韵律词到在当前短语的位置（向后）
	i1	后一韵律词的音节个数
	i2	后一韵律词的词个数
韵律短语层	j1	前一短语的声调类型
	j2	前一短语的音节个数
	j3	前一短语的词个数
	j4	前一短语的韵律词个数
	k1	当前短语的声调类型
	k2	当前短语的音节个数
	k3	当前短语的词个数
	k4	当前短语的韵律词个数
	k5	当前短语到在当前句子的位置（向前）
	k6	当前短语到在当前句子的位置（向后）
	l1	后一短语的声调类型
	l2	后一短语的音节个数
	l3	后一短语的词个数
	l4	后一短语的韵律词个数

语句层	m1	当前句子是否包含有问题声调信息（0 or 1）
	m2	当前句子的音节个数
	m3	当前句子的词个数
	m4	当前句子的韵律词个数
	m5	当前句子的短语个数

攻读学位期间的研究成果

发表论文与申请专利：

[1] Yang H, Hao D, Sun H, et al. Speech enhancement using orthogonal matching pursuit algorithm[C]// IEEE International Conference on Orange Technologies (ICOT), 2014: 101-104.(EI)

[2] 郝东亮, 杨鸿武, 张策, 等. 一种面向汉语统计参数语音合成的生成方法[J]. 计算机工程与应用, 2016, 52 (待刊).

[3] 杨鸿武, 张策, 郝东亮, 等. 一种改进式 TAC 的高精度时间间隔测量系统的实现[J]. 计算机测量与控制, 2015, 23 (12): 4008-4012.

[4] 杨鸿武, 郝东亮. 面向统计参数的语音合成的标注系统. 软件著作权[P]. 登记号: 2014SR154700.

[5] 杨鸿武, 张策, 郝东亮, 等. 一种基于改进式 TAC 的高精度时间间隔测量仪. 实用新型专利[P]. 专利号: ZL201520742600.5.

[6] 杨鸿武, 张策, 郝东亮, 等. 藏语 TTVS 系统的实现方法.发明专利[P]. 公告号: CN105390133A.

参与的科研项目：

[1] 国家自然科学基金项目“汉藏双语个性化语音合成中的语言建模的研究”

[2] 甘肃省杰出青年基金计划项目“汉藏双语跨语言语音合成的研究”

致　谢

本论文是在导师杨鸿武教授的悉心指导下完成的，在论文完成之际，衷心的感谢杨老师的严格要求和耐心指导。研究生期间，在杨老师带领的语音实验室团队中学会了很多，也收获了很多，杨老师对学术严谨的态度深深影响着我，杨老师学识渊博，知识广泛，在学术上不断探索，对每个细节都严抓不放，这种严谨的做事和科研态度对我受益匪浅。除此之外，杨老师随和和谦逊的做人态度也对我有很大的影响。在这里，衷心的感谢杨老师为大家提供的良好的实验室学习条件，以及为大家耐心指导科研问题，特此向恩师表示崇高的敬意和衷心的感谢！

感谢物电学院学院电子系的各位老师，是他们将渊博的知识细心的传授于我，让我学到了完善的理论知识，对研究生所学的课程有了深刻的认识。感谢实验室已毕业的师兄师姐，借助于他们之前的研究成果使我对自己所研究的东西有了很好的认识。感谢实验室一起学习奋斗的同窗们，感谢徐世鹏师兄对我的帮助。

我衷心感谢我的父母和我的亲戚和朋友，对我的支持和鼓励！最后，再次感谢所有对我有过帮助和支持的老师，同学以及亲友！