中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

情感计算
–情感语音合成

中国科学院自动化研究所

刘斌
liubin@nlpr.ia.ac.cn

目录

- 背景及意义
- 研究现状
- 文语转换系统（TTS）
- 语音情感分析
- 情感语音合成系统
- 展望

目录

背景

- 语音不但能够表达出内容信息，而且还能表达出情感、态度以及说话人特性等方面的信息
- 语音合成涉及声学、心理学、语音学、语言学、信号处理、机器学习等多个学科，解决如何将文字转换成语音的问题
- 跟声音回放的差异在于，语音合成能够在任意时刻将任意文字信息转换成语音，让机器和人一样具有说话的能力

背景

- 语音合成包括两个方面，“说了什么”即文本内容和“怎么说的”即合成语音的态度和情感状态
- 现有语音合成基本解决了朗读风格语音合成的可懂度和自然度问题，但是缺乏不同年龄、不同风格、不同语言的表现力
- 具有情感表现力的语音合成日益成为语音合成领域的研究热点

意义

- 情感语音合成技术具有广泛应用
  - 有声文档：情绪化的诗词朗诵、说评书、讲故事
  - 游戏娱乐：根据游戏场景安排输出语音的语气和蕴含的感情色彩
  - 人机交互：根据对话对象的情感状态来调整系统的语音输出

目录

研究现状

- Cahn开发了名为Affect Editor的情感语音合成器，能够合成各种情感色彩的语音
- 英国Bournemouth大学提出了RP-PSOLA的情感语音合成方法，可以通过选择接近给定目标基频的语音片段，将语音单元拼接，合成情感语音
- IBM构建了一种标注语言（Emotional Markup Language），对语音合成系统进行控制以产生恐惧、开心和开心到想哭的语气
- 百度人工智能，语音合成“情感化”

应用

- 百度运用情感语音合成技术还原张国荣声音视频

情感语音合成

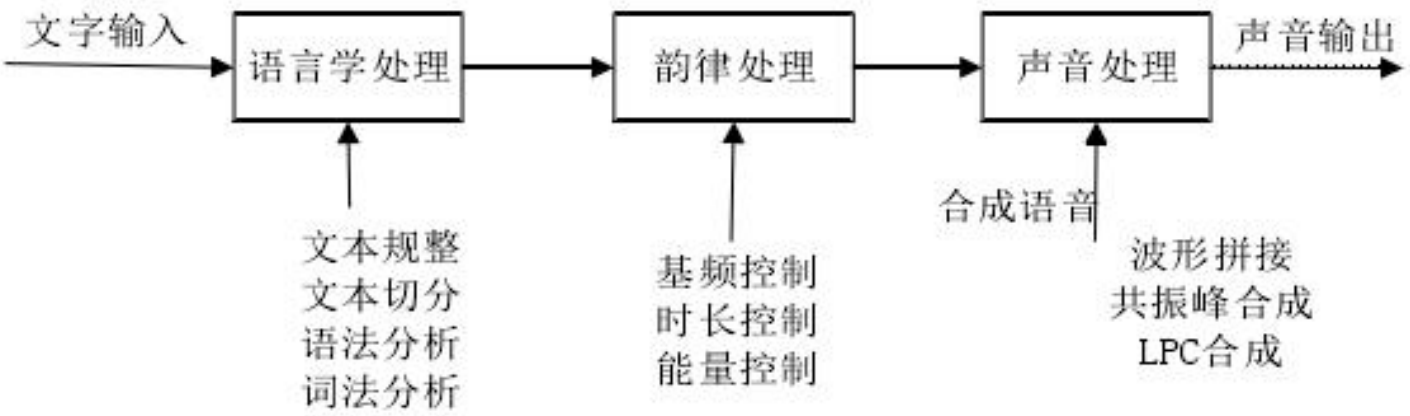
- 情感语音合成
  - 首先是一个语音合成问题
  - 然后是结合情感的语音合成

目录

- 背景及意义
- 研究现状
- 文语转换系统（TTS）
- 语音情感分析
- 情感语音合成系统
- 展望

文语转换系统（TTS）

- 文语转换系统：将文本转换成清晰、自然的语音
  - 文本处理：分词、字音转换，一整套的韵律控制规则
  - 声学处理：语音合成技术

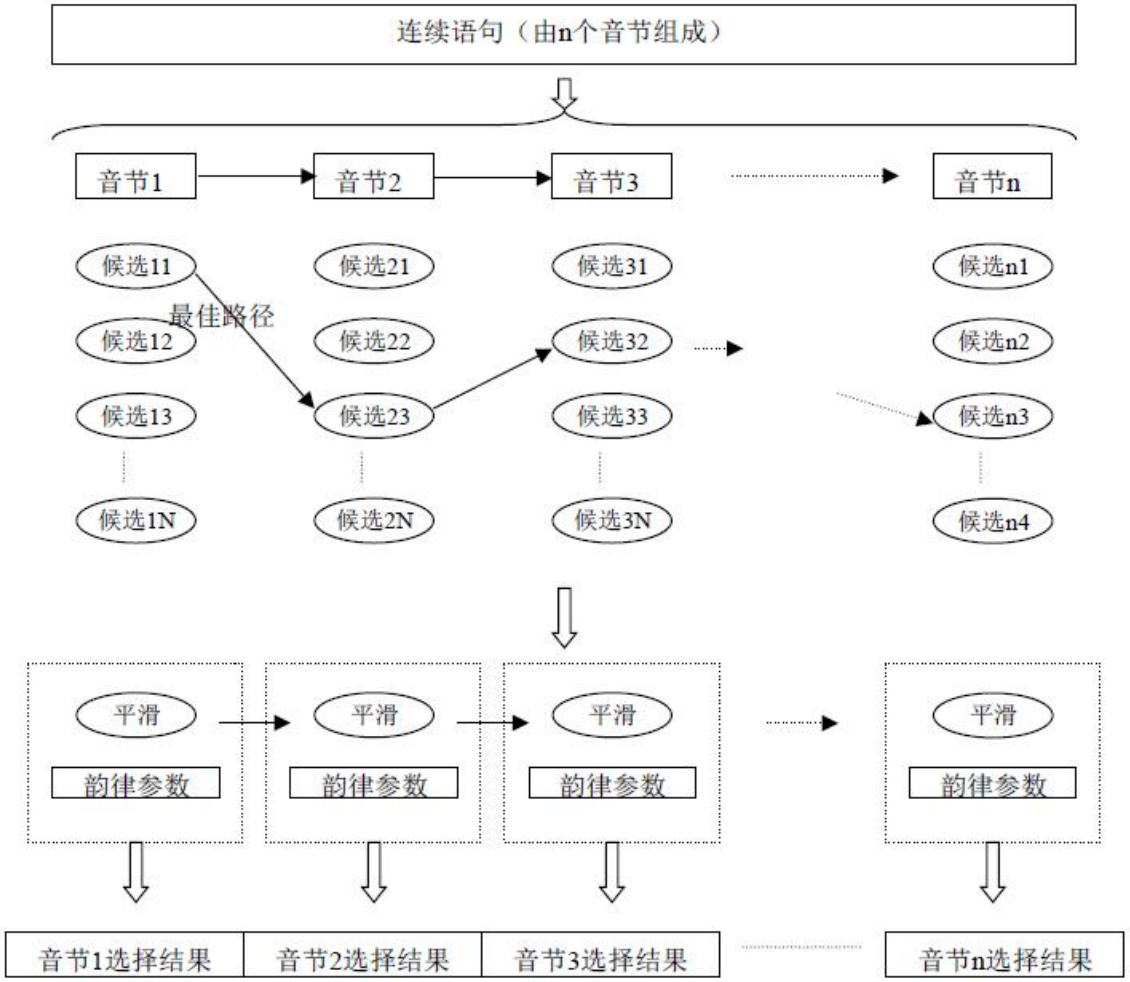


文语转换系统（TTS）

- 波形拼接的语音合成
  - 20世纪80现代，基音同步叠加（PSOLA）合成技术为基于波形拼接的合成技术注入了新的活力。
  - 传统的波形拼接技术最大的问题在于拼接处的不连贯性使得合成语音的自然度大幅下降，而PSOLA根据上下文的要求对拼接单元进行韵律特征的调整，既保留了原始语音的音段特征，又能满足拼接单元的韵律要求。
  - 缺点是对基频提取的准确性要求高，简单的拼接会对频谱带来不利影响，需要数据多

文语转换系统（TTS）

- 波形拼接语音合成系统-连续语句的基元选取方法

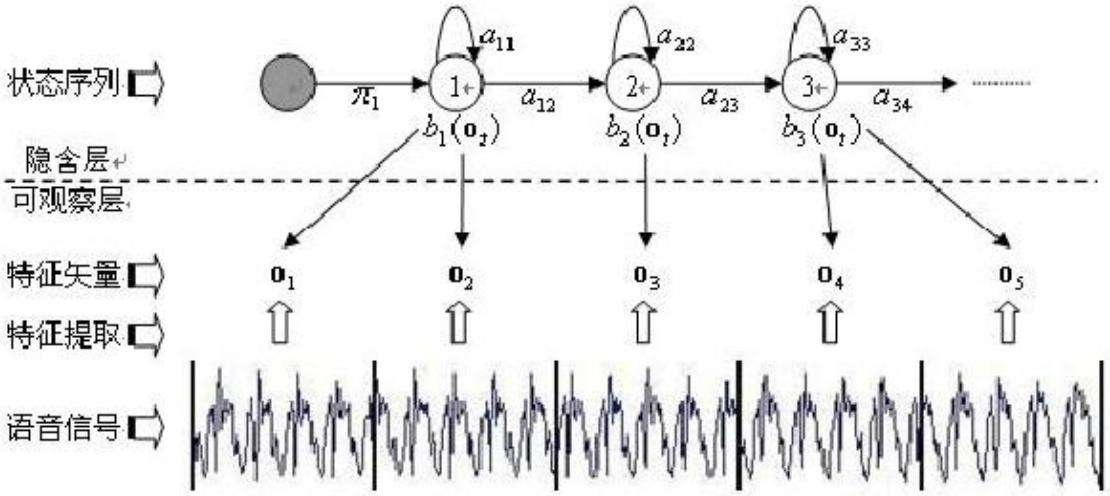


文语转换系统（TTS）

- 发展历程：基于统计建模的语音合成
  - 随着语音信号处理技术的发展以及统计学、模式识别技术的引入，基于统计建模的语音合成方法逐渐发展；
  - 该方法需要对声学参数和标注信息进行统计声学建模。相比于传统的波形拼接技术，具有很少需要人工干预的优点，合成效果稳定。
  - 基于隐马尔可夫模型的语音合成系统HTS应用广泛。近来，深度神经网络方法也取得了巨大的进步。

文语转换系统（TTS）

- 基于隐马尔可夫的语音合成系统
  - 隐马尔可夫一个随机状态描述状态的转移 ， 与语音中的声学参数的变化具有相似性；另一个随机过程描述状态和观察值之间的一个对应关系。HMM很符合人类的语音产生机理。

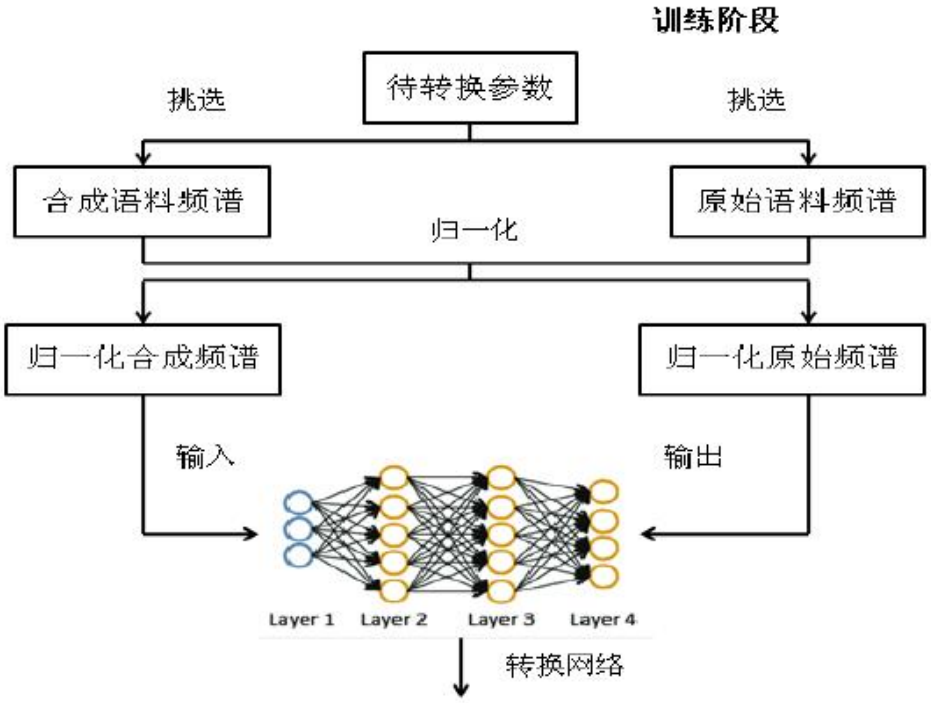


系统小，可以快速构建，稳定性较好

受限于声码器技术，合成音质不高
HMM建模过于平均，韵律表现较差

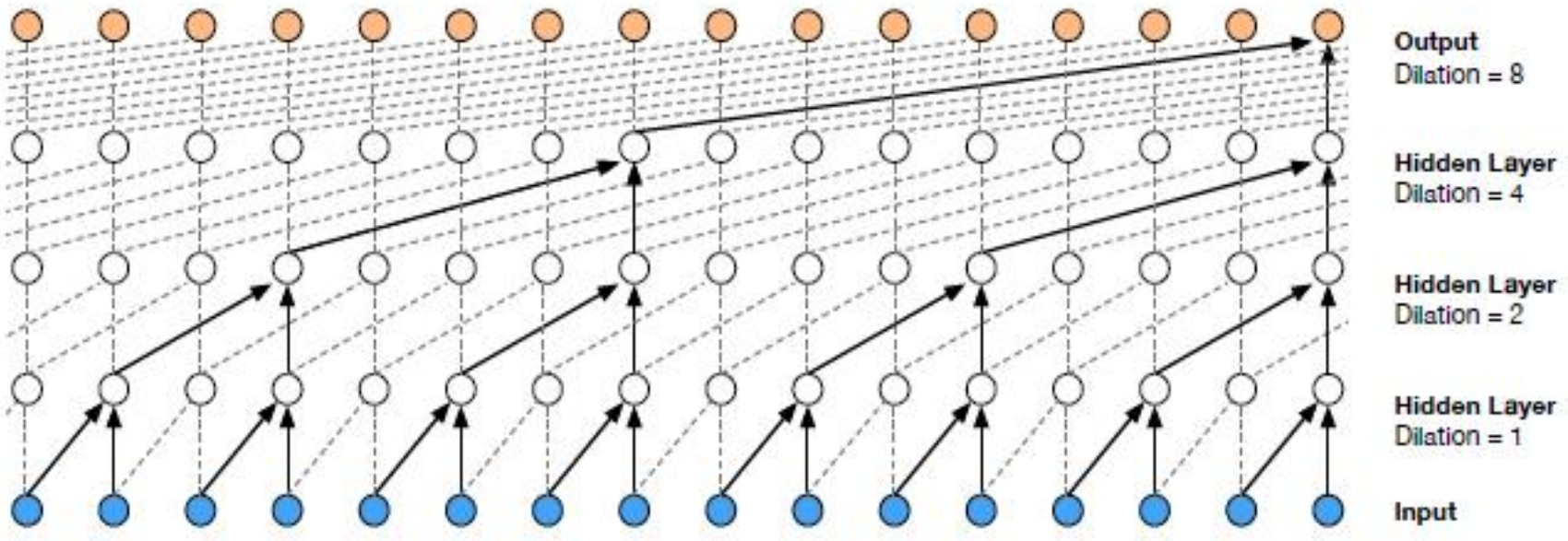
文语转换系统（TTS）

- 基于深度神经网络的语音合成
  - 首先从合成语音和原始语料中选取相同模型单元的平行语料的频谱参数，经过时间对齐后统一进行归一化处理，分别作为深度神经网络的输入参数和输出参数进行学习



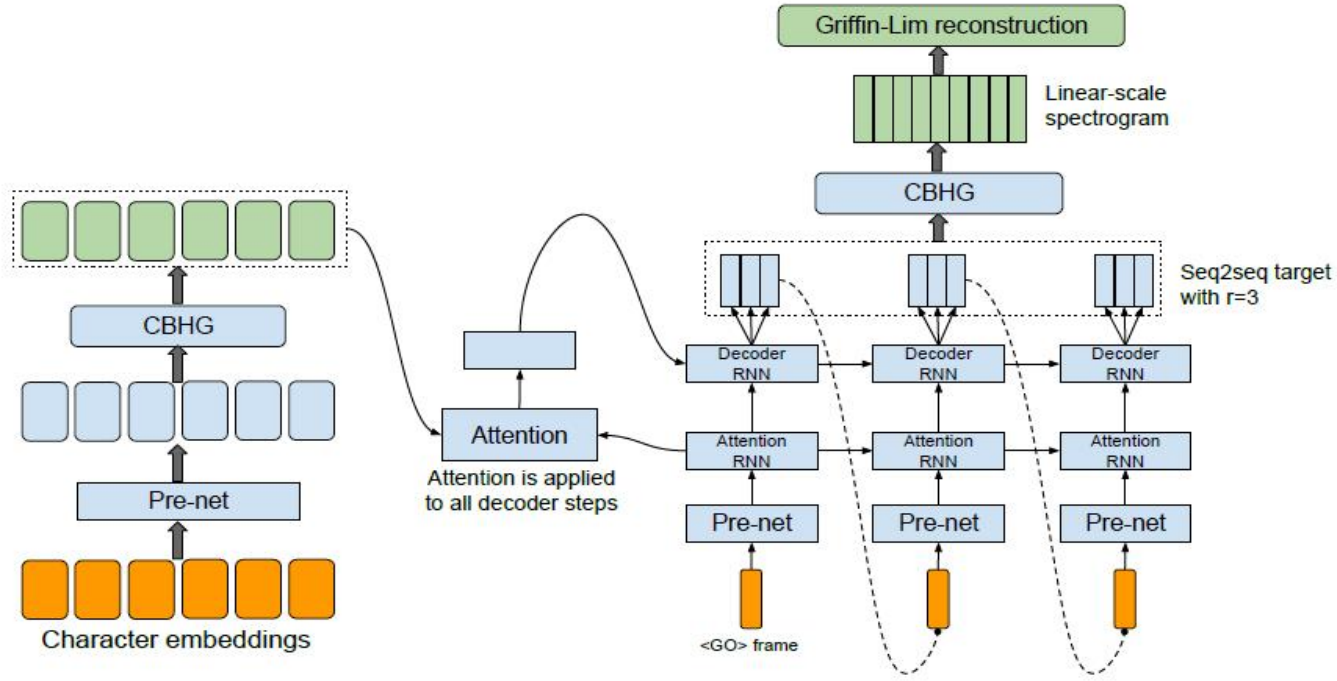
文语转换系统（TTS）

- 端到端的语音合成技术：Wavenet
  - 对时域采样点进行建模，当前采样点是依靠之前若干个采样点和当前采样点的全局和局部文本、韵律风格特征所预测。为了扩大模型的感受野（同等的模型参数的情况大更多的建模采样点数），采用了加宽的卷积结构（dilated causal convolutions）



文语转换系统（TTS）

- 端到端的语音合成技术：Tacotron
  - 文本的字母作为模型的输入，输出频谱参数，最终通过Griffin-Lim算法生成最终的重构语音。Sep2sep的编码器-解码器模型，核心是寻找可以用来重构语音的中间表征特征。



文语转换系统（TTS）

- 语音合成系统评测方法（客观测评）
  - 语音的韵律参数：基频、时长和音强受上下文的影响较大，选用这几种参数作为评测对象
  - 语音的频谱参数：频谱表示声音文件的音色，音色很大程度上影响一段语音的自然度好坏，比如Mel频率的概念就是人耳的音调感知能力

文语转换系统（TTS）

语音合成系统评测方法（主观测评）

主观印象打分（MOS）

级别 MOS 值	MOS 值	用户满意度
优	4.0～5.0	很好，听得清楚，延迟很小，交流流畅。
良	3.5～4.0	稍差，听得清楚，延迟小，交流欠缺顺畅，有点杂音。
中	3.0～3.5	还可以，听不太清，有一定延迟，可以交流。
差	1.5～3.0	勉强，听不太清，延迟较大，交流重复多次。劣 0～1.5
劣	0～1.5	极差，听不懂，延迟大，交流不通畅。

两两对比测试或偏好度测试：将同一段文本的两种待比较系统合成的语音作为一组一次又听音人听，给出优劣评价，最终评价优多的那个系统性能更好

目录

- 背景及意义
- 研究现状
- 文语转换系统（TTS）
- 语音情感分析
- 情感语音合成系统
- 展望

情感语音感知

- 情感相关参数
  - 韵律特征
  - 谱相关特征
  - 音质特征

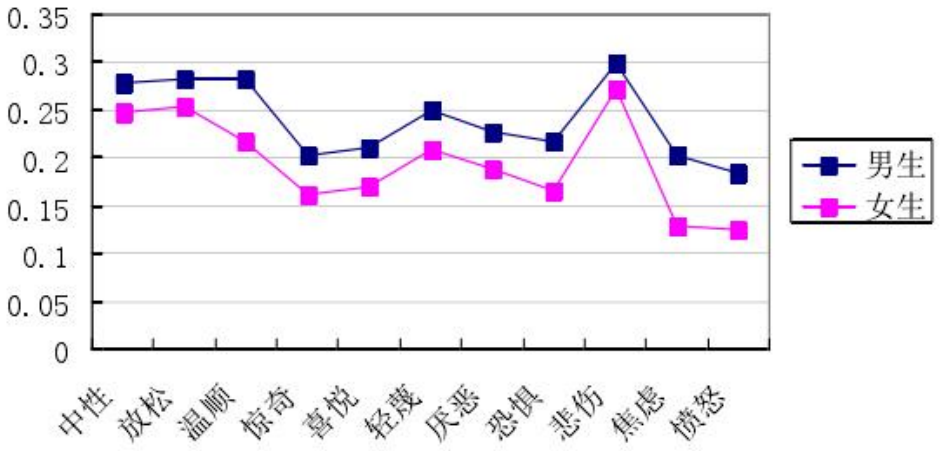
语音情感特征

- 语音情感特征种类
  - 韵律特征：最主要的语音情感特征，如语速、音量和音调等，例如发怒时都会增加；振幅、基音频率，持续时间；
  - 音质特征：音频抖动（Jitter）和振幅抖动（shimmer），谐波噪声率，共振峰；
  - 频谱特征：MFCC,LPCC。

语音情感特征

语速

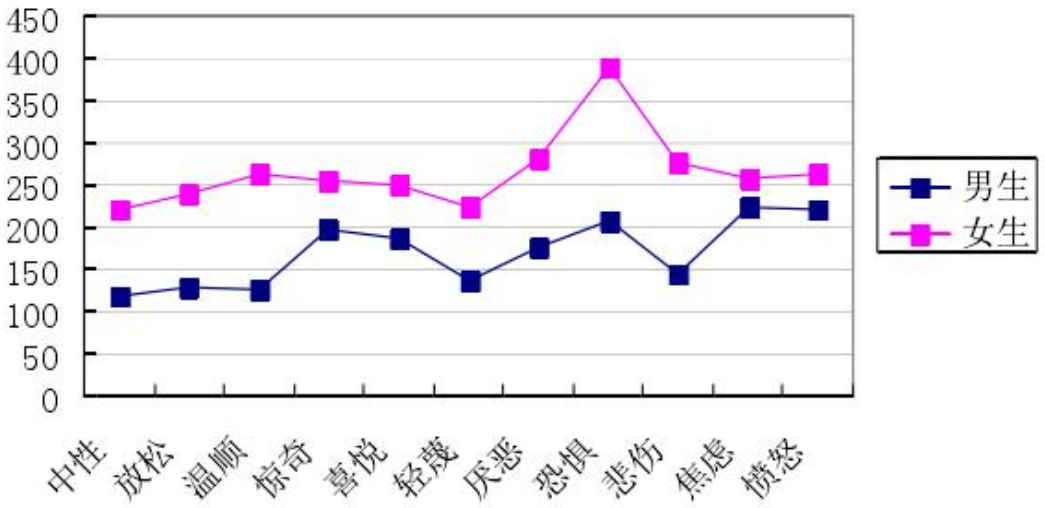
- 语音可以反应出说话者的情绪状态：当人的情绪比较激动的时候，比如处于愤怒状态，语言的表达速度明显加快，相反在人的情绪比较低落时，比如处于悲伤状态，语言的表达速度则明显较慢。



语音情感特征

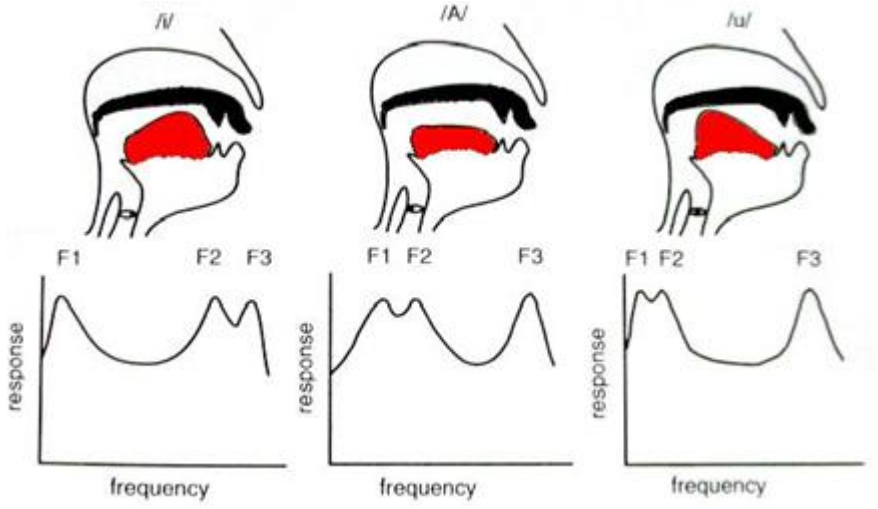
基频

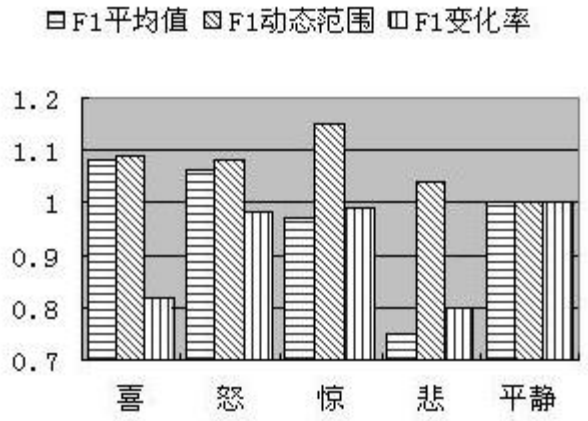
- 基音频率体现出以下规律：处在激动情绪下如愤怒的人所表达出的语音的基频较高，变化范围较大；处于低落情绪如悲伤的人所表达的语音的基频较低，变化范围较小，处于平静情绪下的人所表达出的语音的基频则相对稳定。



语音情感特征

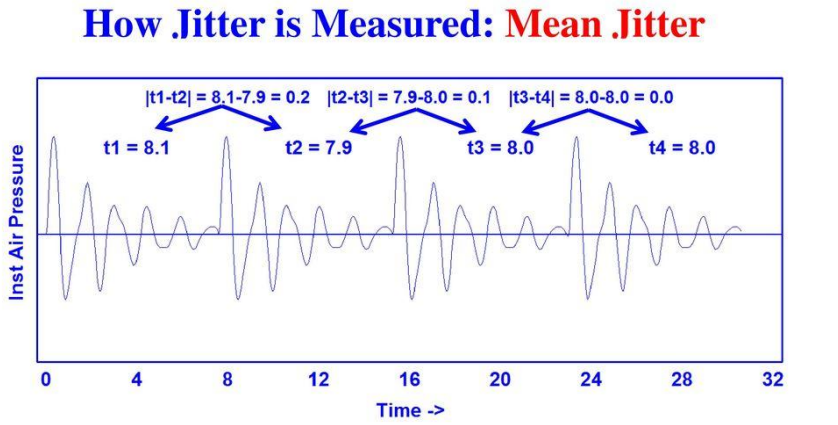
- 共振峰
  - 共振峰是反映声道特性的一个重要参数。不同情感发音的共振峰的位置不同。





语音情感特征

- 基频抖动（Jitter）
  - 焦虑语音会出现“F0抖动”现象。Jitter是基频值的变化程度。
  - F0 Jitter是由生理器官的作用才产生，比如情感的变化会导致声带肌肉紧张度，气流的体积速度，声道表面的坚硬或柔软等发生变化，从而产生基频抖动现象。



语音情感特征

- 线性预测倒谱系数（LPCC）
  - LPCC是基于语音信号为自回归信号的假设，利用线性预测分析获得倒谱系数。
  - 不同情感的发音会使声道有不同的变化，进而引起声道传输函数倒谱的变化。
- Mel频域倒谱系数（MFCC）
  - MFCC考虑了人耳对不同频带的分辨率不同，充分融合了人耳的听觉特性。

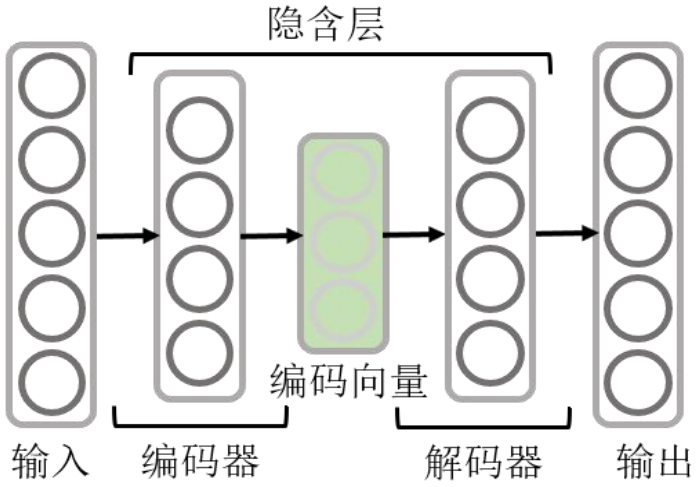
语音情感特征

- 功能性副语言中携带了大量情感信息

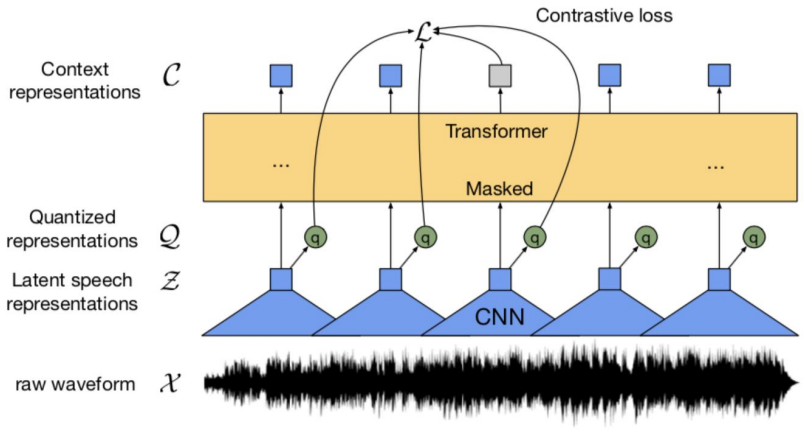
功能性副语言	高兴	伤心	惊讶	生气	害怕	厌恶
笑声	Y	N	N	N	N	N
伤心的哭声	N	Y	N	N	N	N
质疑声	N	N	Y	N	N	N
叫喊声	N	N	N	Y	N	N
害怕的哭声	N	N	N	N	Y	N
叹息声	N	N	N	N	N	Y

语音情感特征

基于深度学习的情感特征



(a) 自编码器



(b) wav2vec

目录

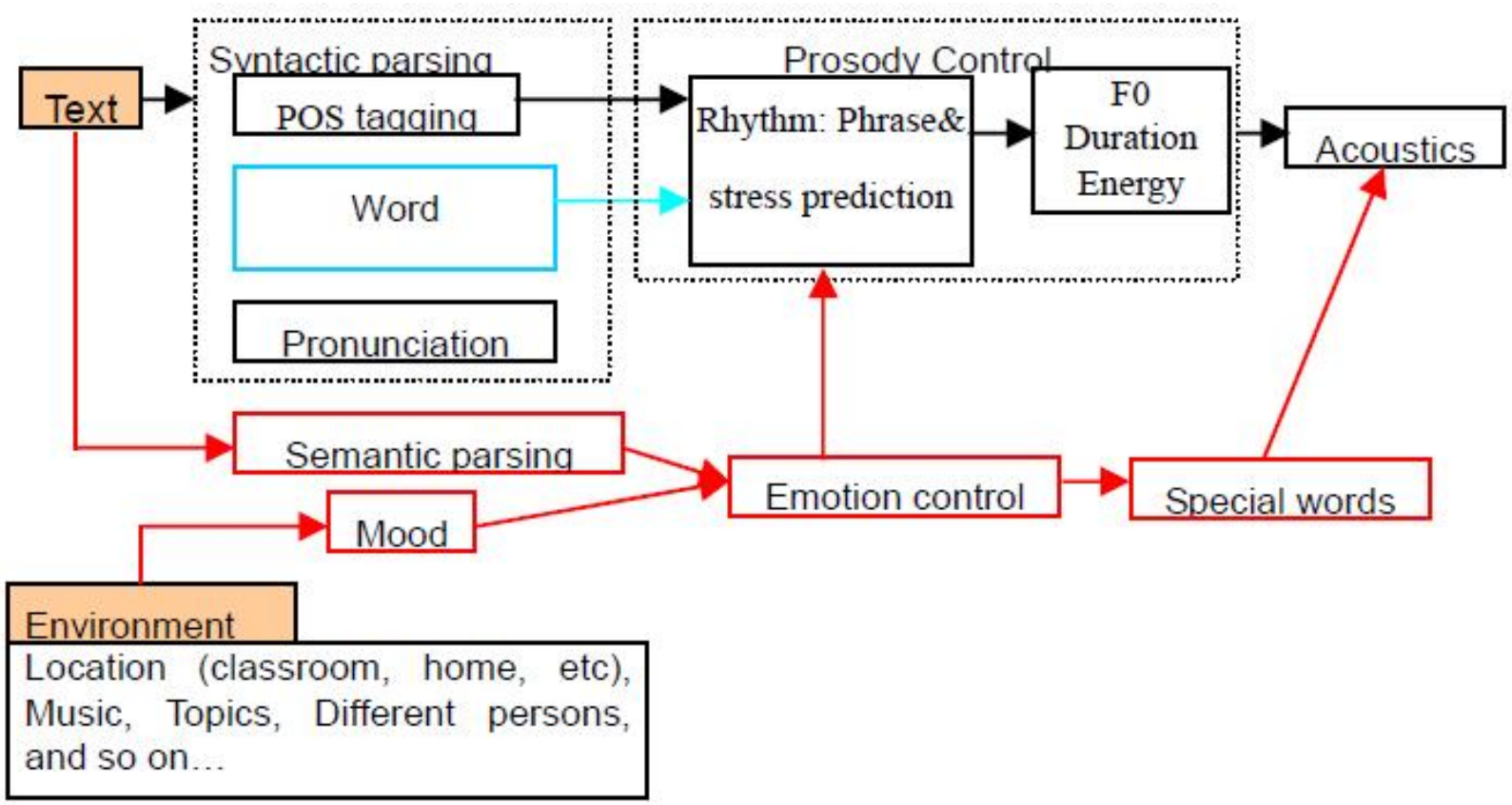
- 背景及意义
- 研究现状
- 文语转换系统（TTS）
- 语音情感分析
- 情感语音合成系统
- 展望

情感语音关注成分

- 情感焦点词
  - 情感关键词：情感焦点在通常情况下，由情感关键词驱动，多出现在情景对话和具有剧烈变化的情感状态中。
  - 比如：我非常生气。短语“非常生气”表示了句子的关键的情感状态，并且在愤怒的情感时会得到有力地加强。其它的一些词语，如：“不好”，“很”，“非常”等等也会达到同样的效果。

基于场景驱动的情感语音合成

- 情感发音，不能离开外界因素对人发音的影响。完整的情感语音合成系统需要引入了环境感知和情感控制模型。



情感语音关注成分

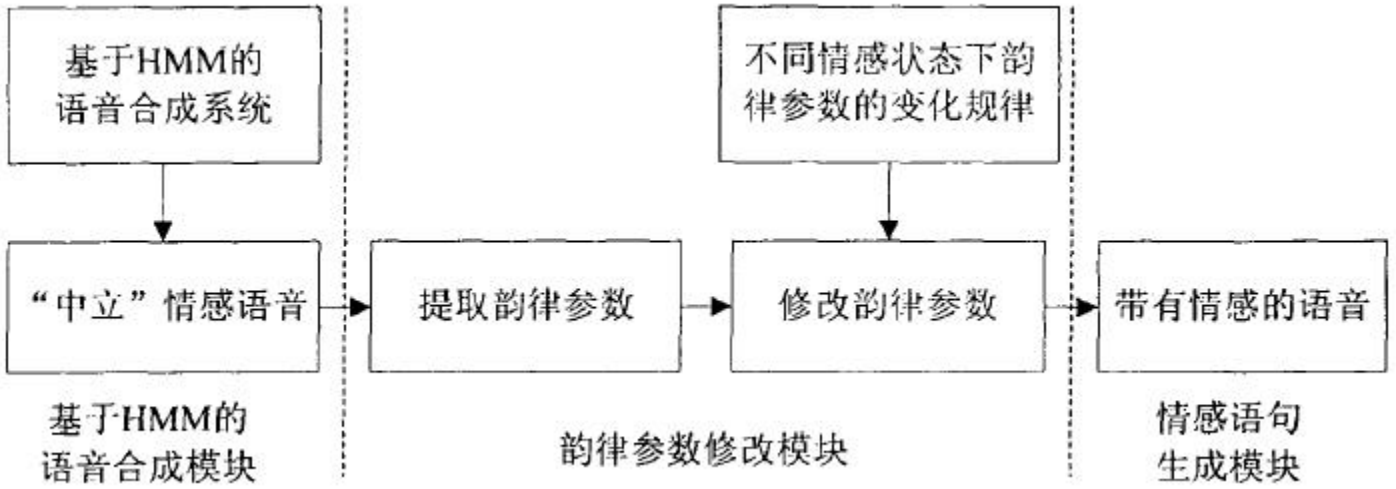
基频

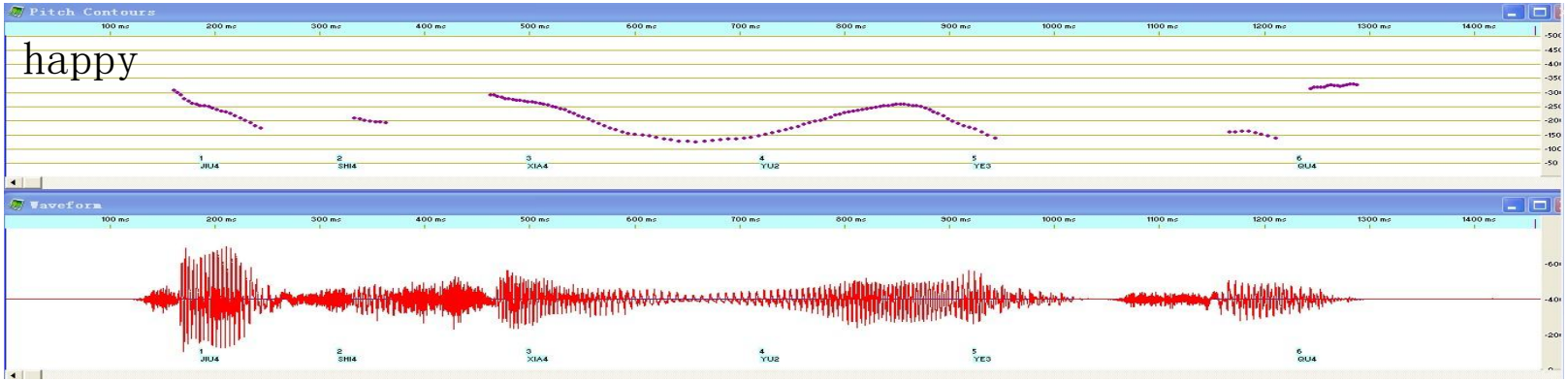
- 随时间变化的，基频的变化形成了语调。语调是说话人心理、情感和态度的反映

语调类型	描述
平语调	没有高低升降的字调变化。居中个音节的声调基本上是原状，只是句尾的音高是趋降的。可以表达陈述、说明。
升语调	末尾语调上扬，音高上升。调域扩大，调长加长。升语调可用于表达命令、疑惑、惊奇等
降语调	降语调的末音节下降，可以表示说话人情绪不高，感叹
曲语调	曲语调是一种混合组成语调，主要用于表达较复杂的感情和语义

基于韵律转换的情感语音合成

- 基频能反映情绪状态波动=> 基频转换

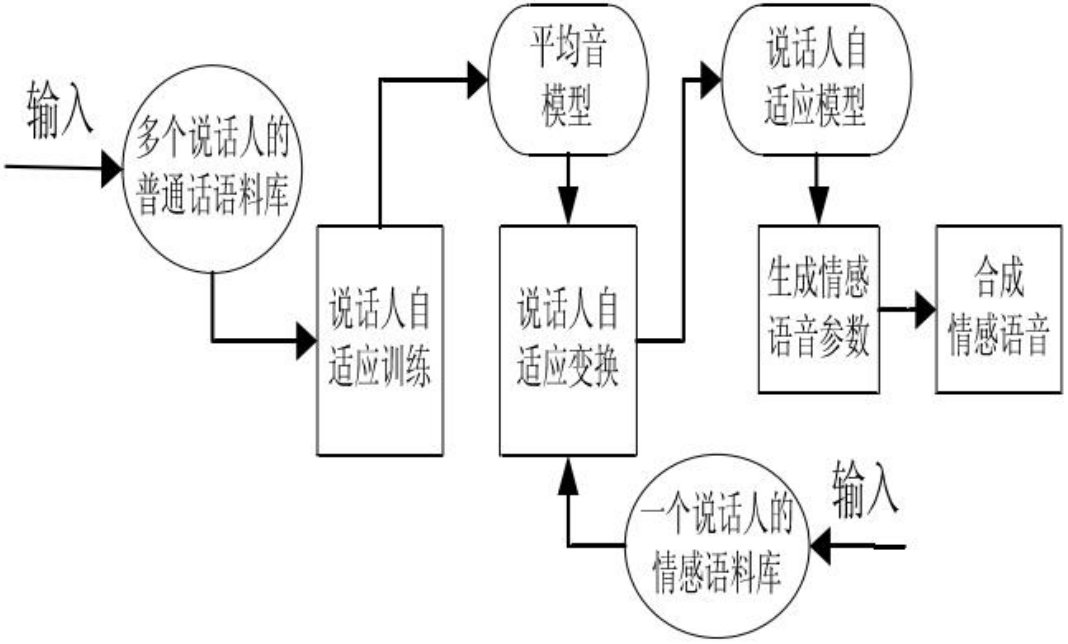




基于说话人自适应的情感语音合成

系统框架

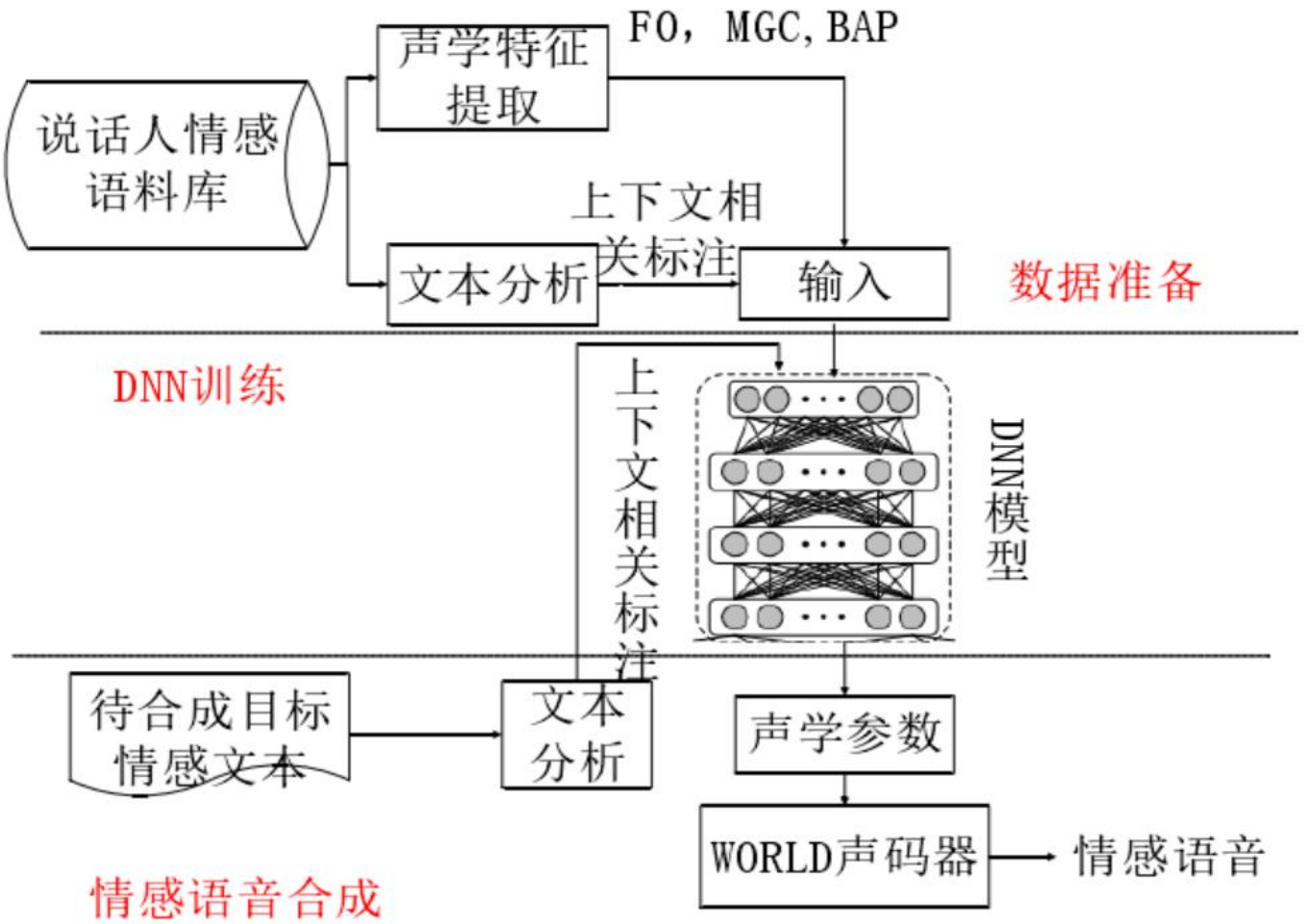
- 选取一个多说话人的普通话大语料库和一个说话人的情感小语料作为训练语料，通过说话人自适应训练获得一个混合语言平均模型
- 采用说话人自适应变换方法，利用情感特定说话人的训练语料，获得说话人相关的情感语音模型，最后合成情感语音。



基于DNN的情感语音合成

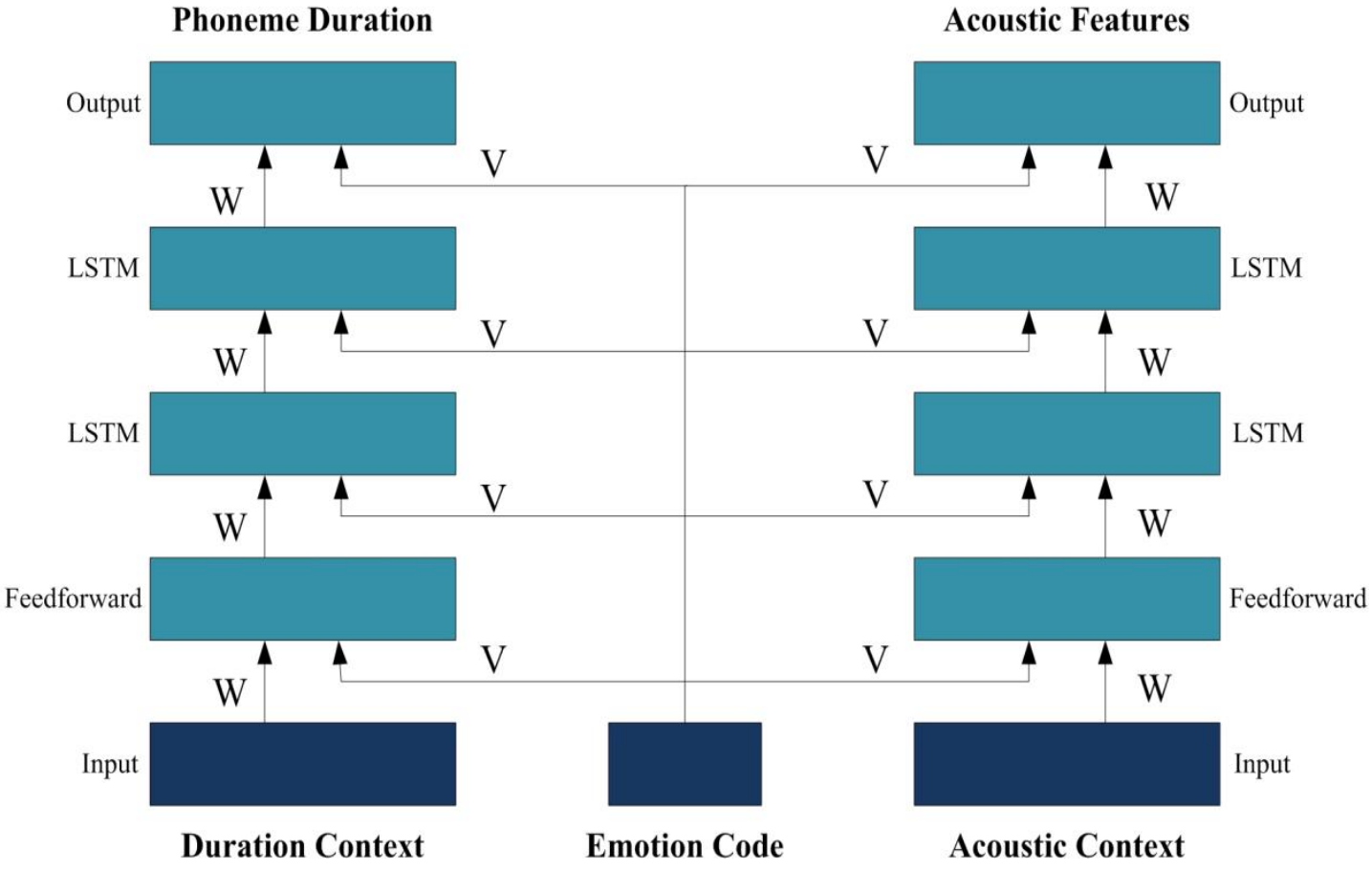
系统框架

- 将情感向量引入声学模型，从而预测出带有情感内涵特征参数。



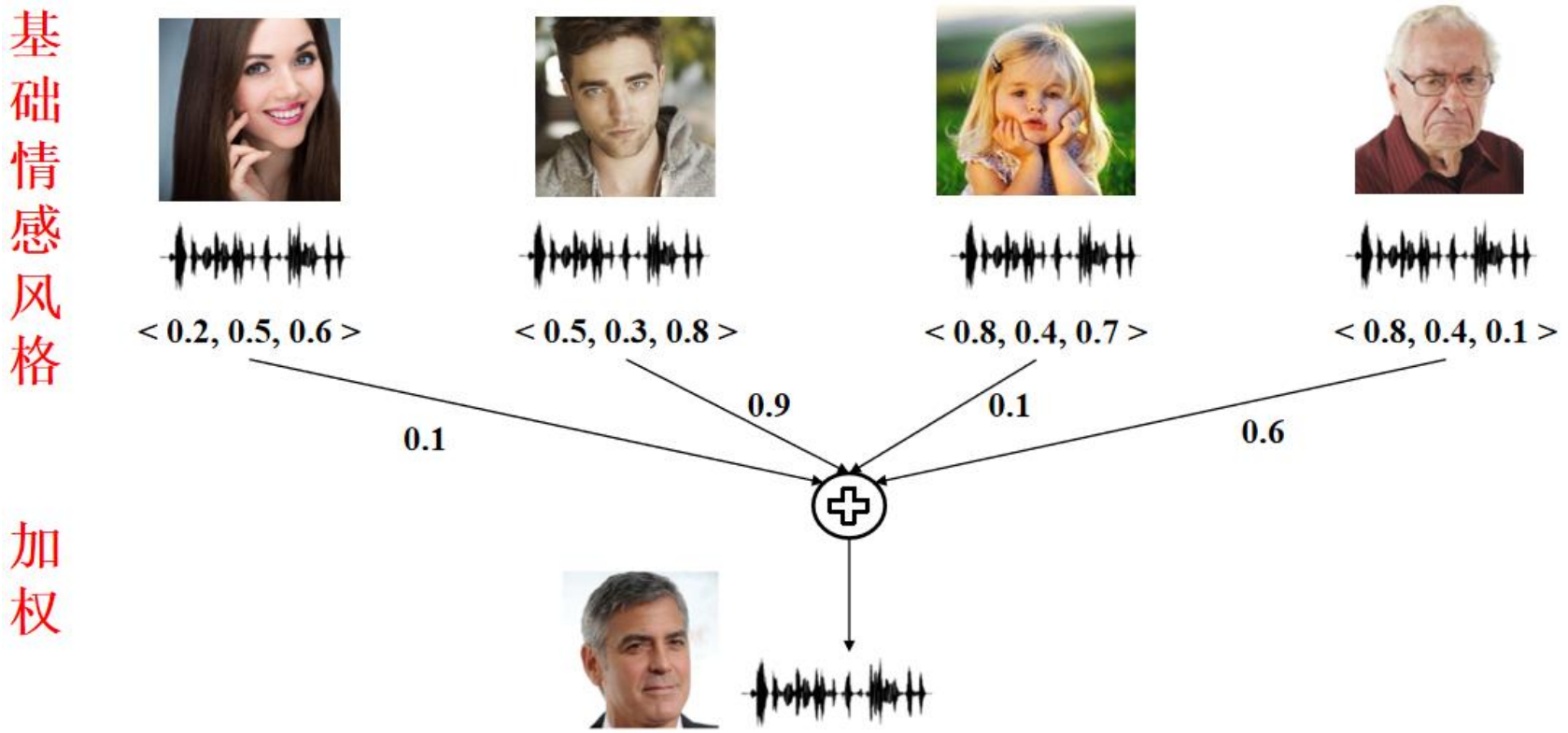
基于RNN的情感语音合成

- 系统框架
  - 将情感向量引入时长模型以及声学模型。



基于情感风格加权的情感语音合成

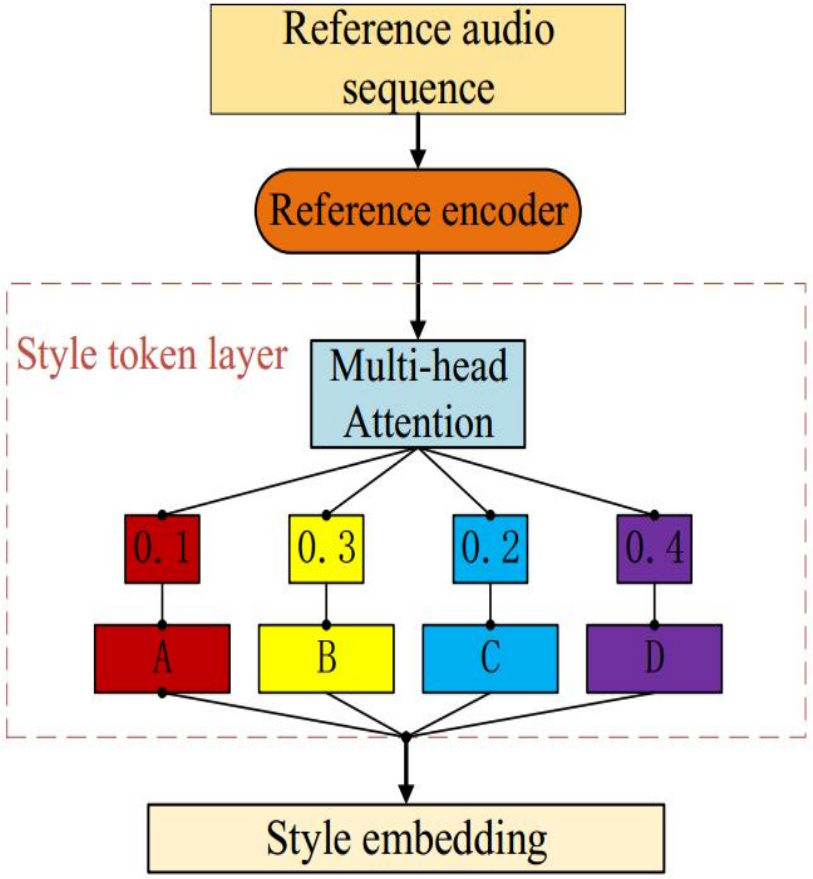
- 系统框架
  - 将不同发音人的基础情感风格进行动态加权



基于情感风格加权的情感语音合成

系统框架

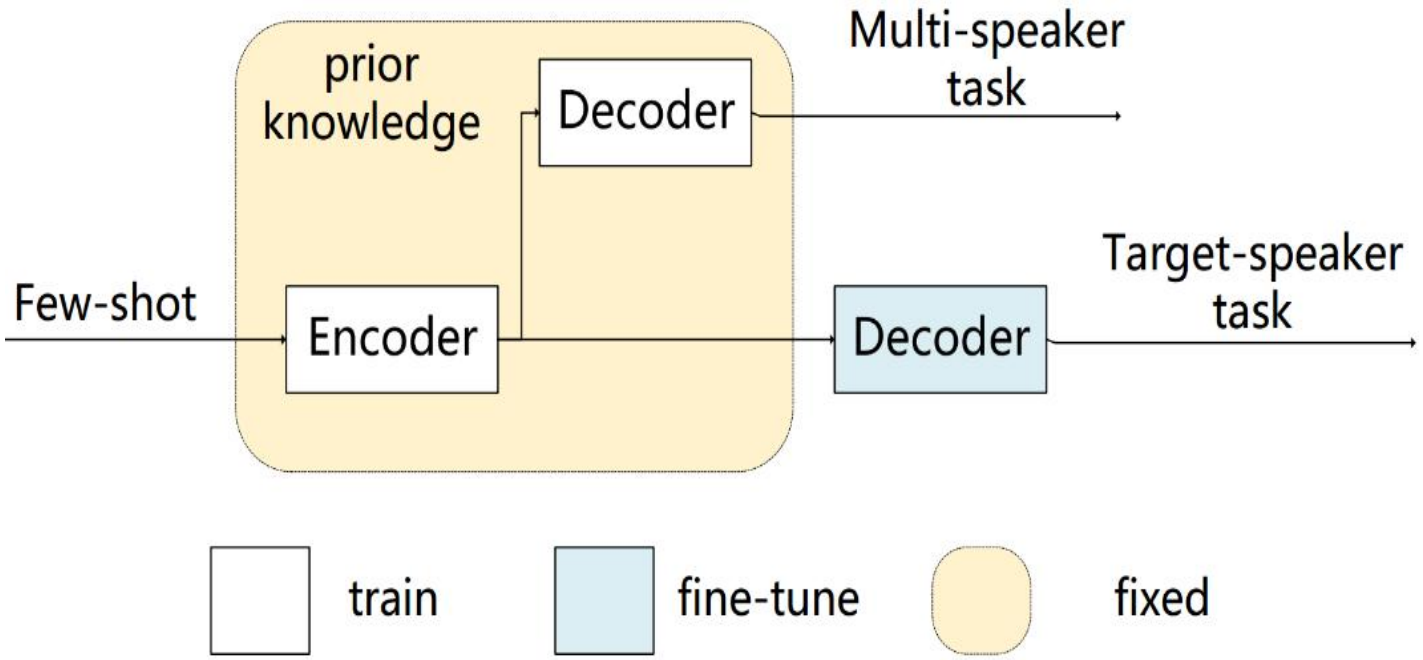
基于注意力机制的情感风格特征抽取



情感风格相关参数的快速自适应

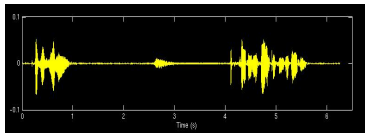
系统框架

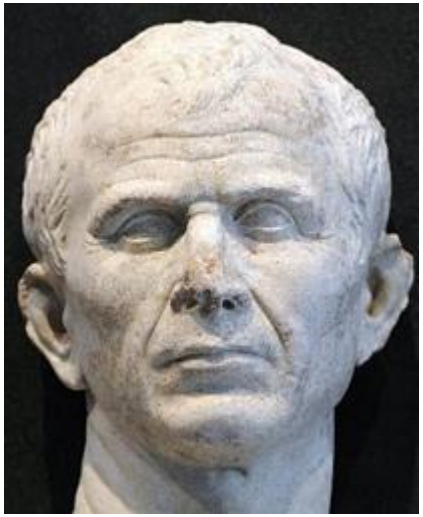
极低资源条件下的情感语音快速自适应



具有情感表现力的可视语音合成

- 可视语音（也称视觉语音）是指话语者发音过程中可视发音器官（如嘴唇、舌头、下腭、面部肌肉等）的动态变化过程
- 具有情感表现力的可视语音是指合成的可视语音能够同时反映话语者说话过程中所具有的情感状态。
- 声音的可视化对语音内容的正确理解（尤其在有噪声环境下或者听者有听力障碍的情况下）是至关重要的。





具有情感表现力的可视语音合成

- 基于关键帧差值方法
  - 关键帧是指能够反映人脸重要姿态、动作及表情的单一图像帧。
  - 关键帧插值法是指根据输入的文本或语音，通过规则或者映射关系在图像样本库中找出所需的带有情感信息的关键帧
  - 在相邻关键帧间进行插值，从而得到具有平滑过渡效果的情感可视语音的过程。

具有情感表现力的可视语音合成

- 基于图像序列拼接方法
  - 首先建立包含大量说话时人脸图像的样本库，建立每个发音与图像序列之间的对应关系；
  - 合成时根据发音或文本内容从库中选择出合适的图像序列，并进行拼接，从而生成动态连续的可视语音。

目录

- 背景及意义
- 研究现状
- 文语转换系统（TTS）
- 语音情感分析
- 情感语音合成系统
- 展望

总结与展望

- 情感数据库构建：情感语音合成离不开对包含韵律完备的情感语音数据的分析统计，要求拥有大规模、高真实感的情感语音库。目前研究中成规模、标准化的情感语音数据匮乏，迫切需要进一步完善数据。

- 区分行情感特征抽取：情感语音合成中不是单一特征就能完全达到效果，如何综合利用合适的语音特征，制定出有效的情感语音合成规则将会是语音合成中的一个具有挑战性的课题。

- 情感语音合成：目前水平下的合成语音很难体现出高表现力情感，例如在韵律表现上不够灵活，声调变化上相对死板。摆脱平铺直叙，使合成语言更具有表现力依旧是语音合成技术的一大难点。

Q&A