南京大学 NANJING UNIVERSITY

智能科学与技术学院
School of Intelligence Science and Technology

智能语音技术

Intelligent Speech Technology

王帅
准聘副教授

南京大学 NANJING UNIVERSITY
智能科学与技术学院 School of Intelligence Science and Technology

正式开始之前...

王帅

2025.08.25

关于我

王帅

准聘副教授，特聘研究员
博士生导师

南雍楼 西536

shuaiwang@nju.edu.cn
https://shuaiwang-nju.github.io

曾任腾讯高级研究员，领导光子工作室语音技术团队。学术界+工业界复合背景。
与工业界保持紧密合作，研究以解决现实场景中的实际问题为导向。

应用场景

语音是最自然的交互方式

车载交互

智能音箱

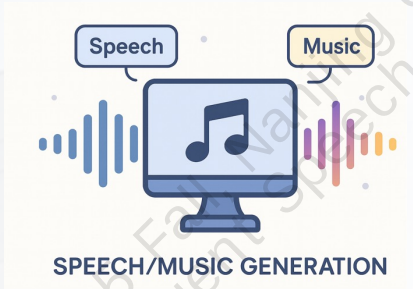
手机助手

研究方向

智能音频处理，结合大模型的音频感知、认知与生成

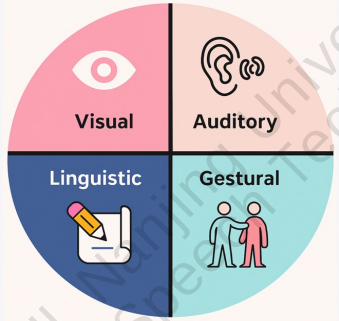
丰富音频信号理解

对人声信号的全面建模
声纹识别、语种识别、情感计算



可控、可信音频生成

高质量语音、音乐生成
隐私保护、音频水印及溯源技术



多模态感知与认知

融合其他模态，展开跨模态研究
模拟人的感知和认知功能

跨学科应用

其他学科下的声波信号
基于音频信号的疾病检测

教学大纲

1. 语音技术初窥 2
2. 语音信号处理基础 1
3. 语音技术细分领域介绍 11
    1. 语音识别 3
    2. 语音合成 3
    3. 语音转换 1
    4. 语音分离 2
    5. 说话人建模 2
4. 大模型时代的语音技术 2
5. 外部专家讲座 2

语音技术初窥 （2次课，4个学时）

1. 声音的基本概念
2. 人的发音与听觉机制
3. 语音技术的发展史
2. 语音信号处理基础
3. 语音技术细分领域介绍
    1. 语音识别
    2. 语音合成
    3. 语音转换
    4. 说话人建模
    5. 语音分离
4. 大模型时代的语音技术

语音信号处理基础（共1次课，2个学时）

1. 发声系统与听觉系统
2. 语音编解码方法
3. 语音特征提取方法

语音识别 （共3次课，6个学时）

1. 统计语音识别的基本概念和原理
2. 高斯混合模型，隐马尔可夫模型
3. 统计语言模型
4. 基于深度学习的声学模型和语言模型
5. 端到端语音识别

语音合成（共3次课，6个学时）

1. 常见生成模型Recap
2. 传统语音合成系统
3. 基于深度学习的语音合成系统
4. 端到端语音合成
5. 歌声合成

语音转换（共1次课，2个学时）

1. 基于多模块的语音转换技术
2. 语音自解耦方法
3. 流式语音转换
4. 歌声转换

说话人建模（共2次课，4个学时）

1. 因子分析模型，**i-vector**
2. 深度学习时代，**d-vector, x-vector, r-vector**
3. 说话人建模在相关任务中的应用及思考

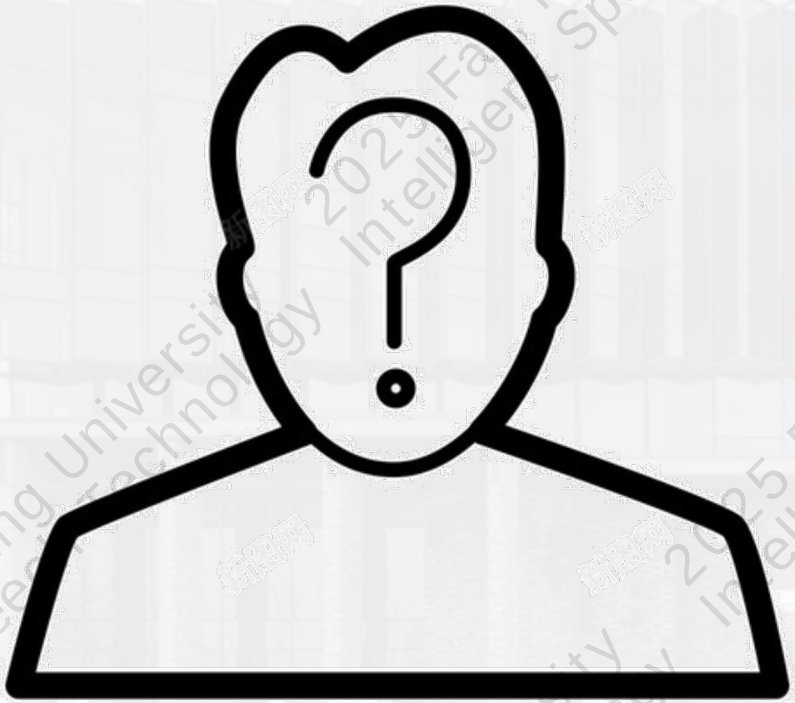
语音分离（共2次课，4个学时）

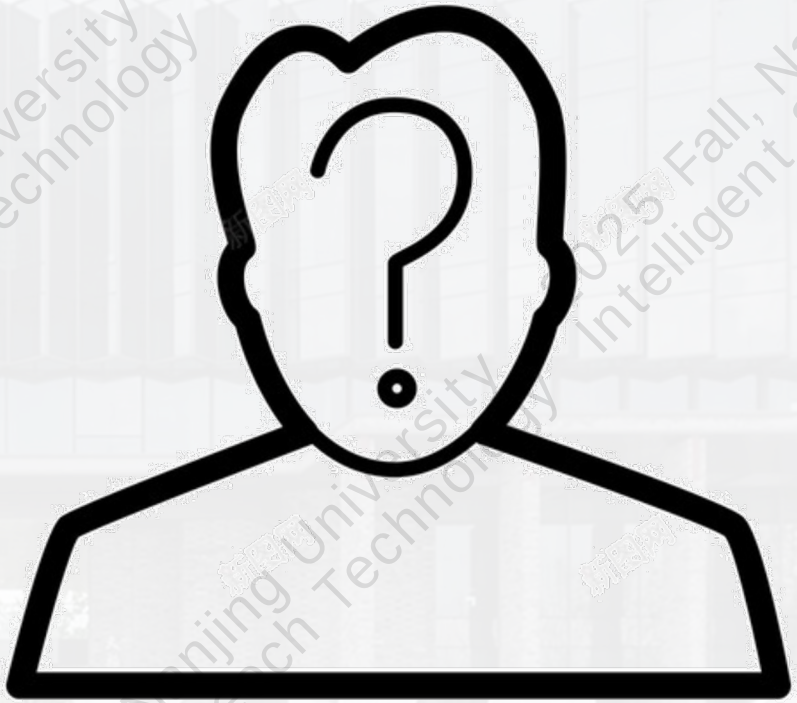
1. 鸡尾酒会问题
2. 传统语音分离方案
3. 基于深度学习的语音分离技术
4. 目标语音提取

基于大模型语音技术（共2次，4个学时）

1. 自监督学习、对比学习方法
2. 大语言模型技术
3. 基于全新范式的语音处理技术
4. 全双工对话系统

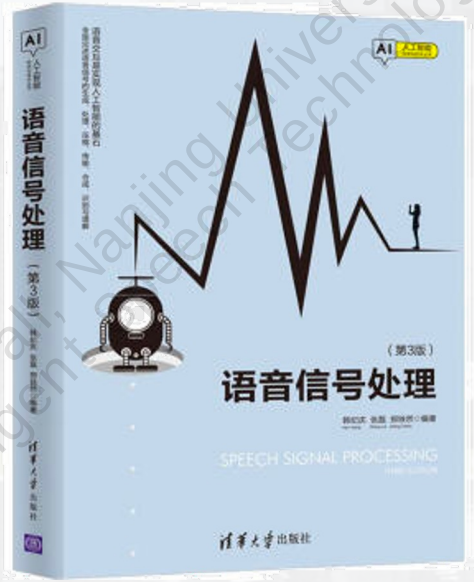
外部专家讲座（共两次课，四个学时）

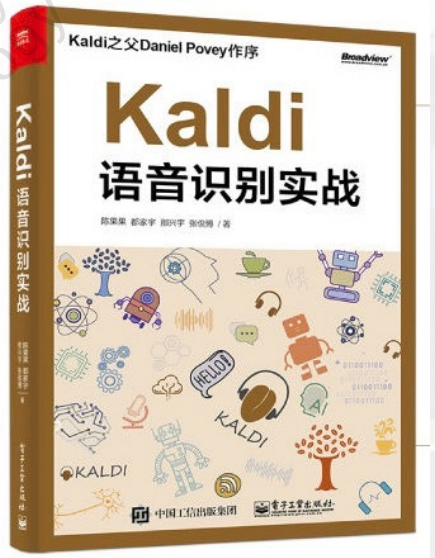




推荐书籍

[1] Jurafsky D, Martin J H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models. 2025 Edition.
[2] https://speechprocessingbook.aalto.fi/index.html
[3] Xuedong Huang, Alex Aceoro, Hsiao-Wuen Hon, Spoken Language Processing: A guide to theory, algorithm, and system development, Prentice Hall, 2011
[4] 韩纪庆、张磊、郑铁然，《语音信号处理》，清华大学出版社
[5] 洪青阳，李琳著，《语音识别：原理与应用》，电子工业出版社





课程考核

1. 出勤 15%（前三周不算，后面每次课 1%）
2. 课程作业 20%（共两次作业，每次10%）
3. 课程项目 65% (规定题目或自选语音相关题目，后者需我确认）
4. 加分项（Bonus) – 10% (上限加到100%）
   - 课程作业形成论文投稿 (Interspeech 26 DDL: 2026.2.25)
   - 贡献开源工具

你将学到什么？

希望本节课可以带给大家

01

人类的发音与听觉机制

02

语音技术的历史，现状与趋势

语音技术的发展历史
世界前沿，技术的发展方向，
面临的挑战

03

智能语音技术的细分任务

了解相关技术形态，输入输出以及基本原理

04

激发对语音技术研究的兴趣

开发你自己的智能语音系统，进一步从事相关研究

你上本门课的目标是什么？

1. 学分 + 2
2. 看着挺有意思的，过来听听
3. 希望以后有机会从事相关研究

你现在的状态是什么

1. 已保研
2. 准备工作/考研
3. 准备申请境外高校

南京大学 NANJING UNIVERSITY
智能科学与技术学院 School of Intelligence Science and Technology

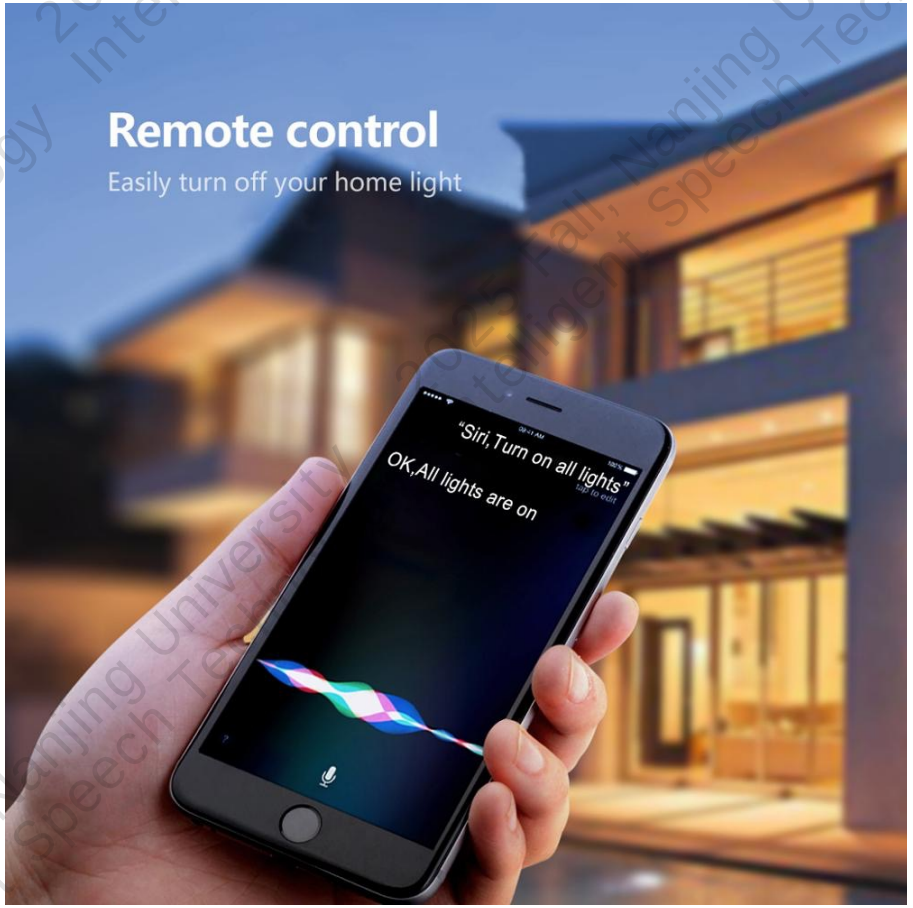
走近智能语音技术

王帅

2025.08.25

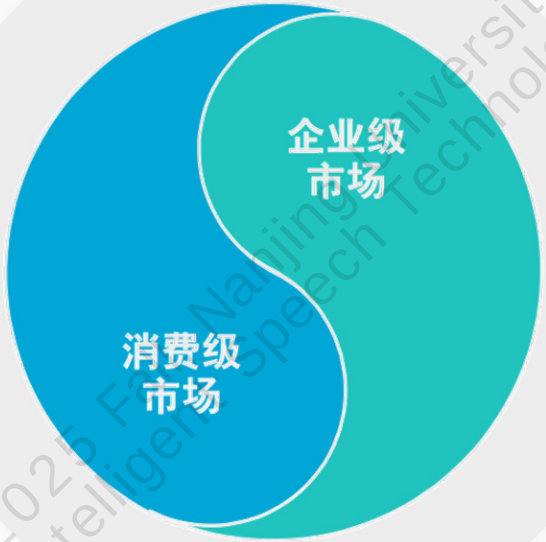
我们身边的智能语音技术

智能家居控制



我们身边的智能语音技术

智能语音技术的细分应用场景



智慧生活
- 手机语音助手
- 智能可穿戴

发展空简约 50亿元

智能家居
- 智能音箱
- 智能家电

发展空简约 503亿元

智慧办公
- 翻译机
- 录音笔

发展空简约 6亿元

智慧驾驶
- 车载语音系统

发展空简约 126亿元

其他

智慧教育
- 智慧课堂
- 学习机

发展空简约 77亿元

智慧医疗
- 电子语音病例

发展空简约 218亿元

智慧电信/电商/金融
- 智能呼叫/客服

发展空简约 419亿元

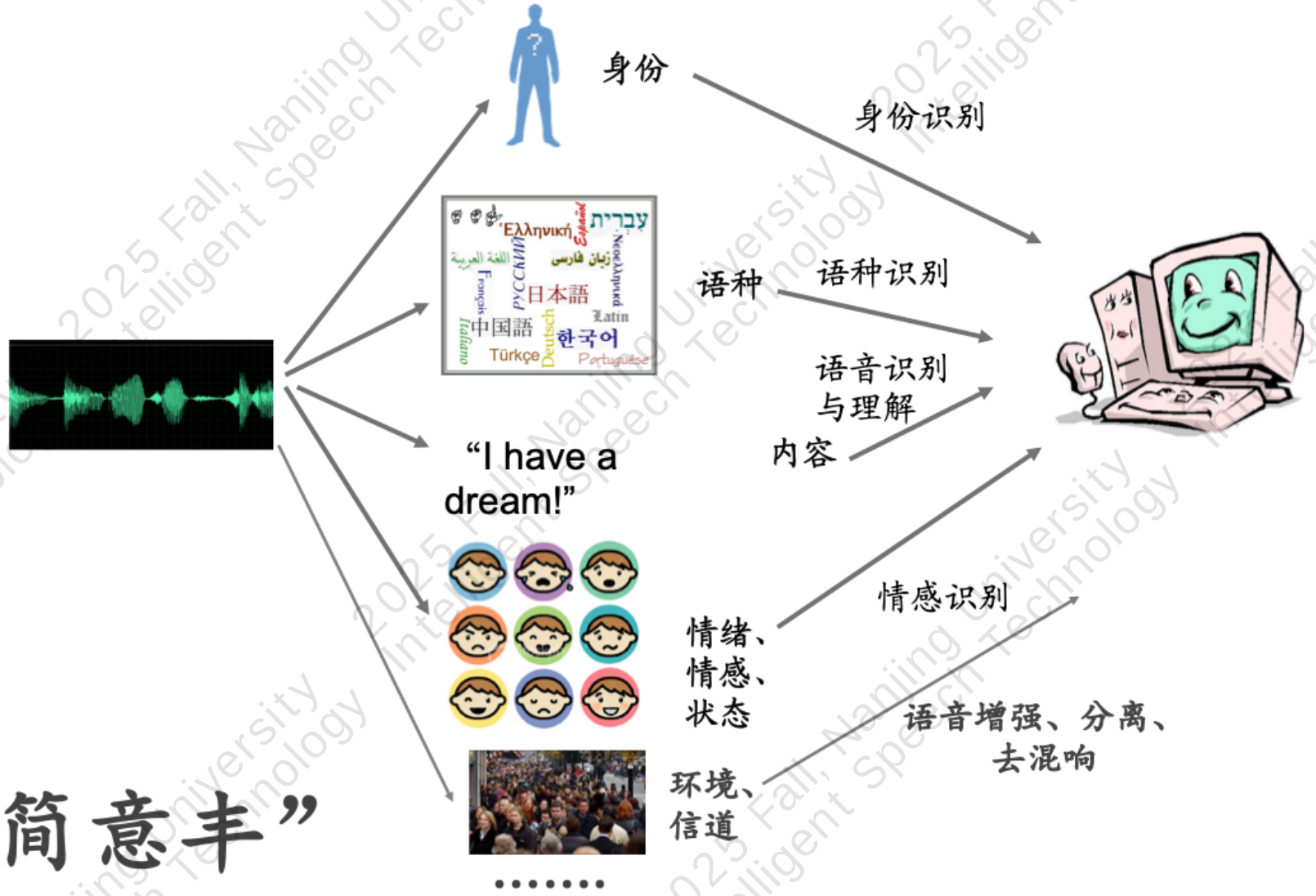
其他

资料来源：iResearch，华西证券研究所，德勤研究

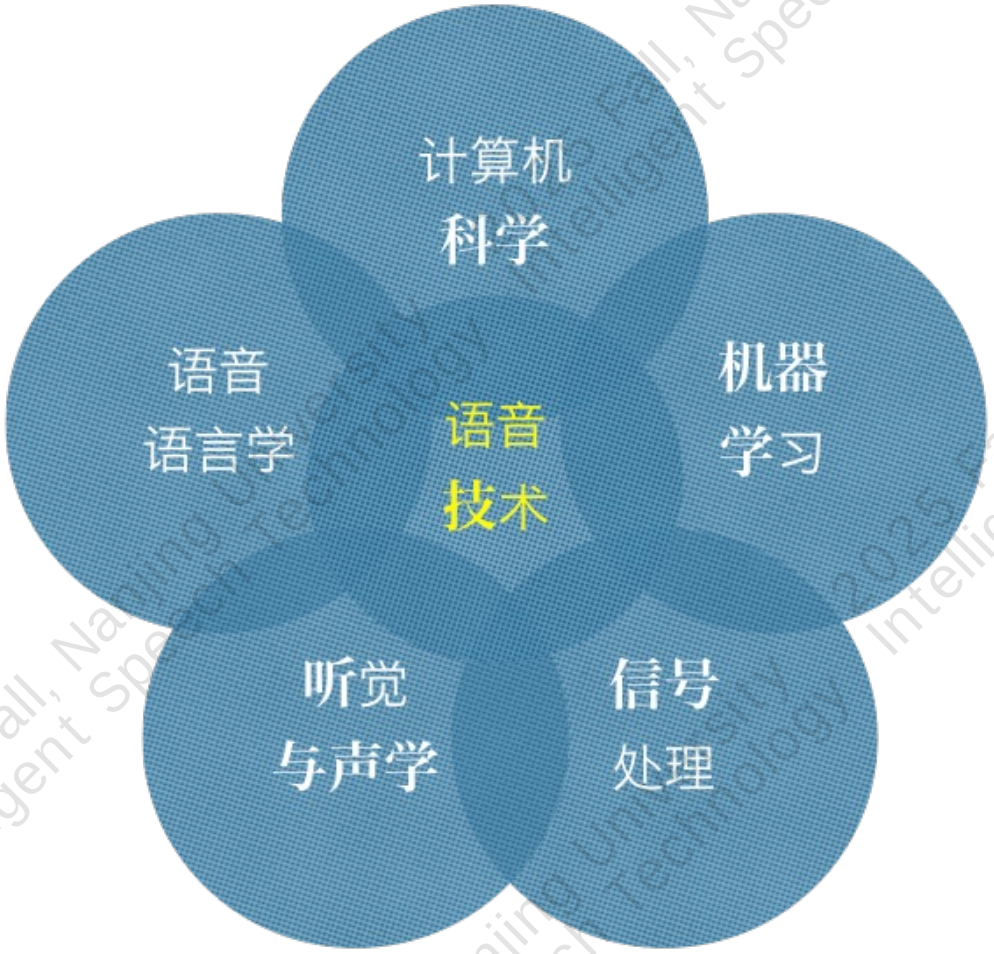
目录

CONTENTS

智能科学与技术学院
School of Intelligence Science and Technology


身份
身份识别
语种
语种识别
语音识别
与理解
内容
“I have a dream!”
情绪、
情感、
状态
情感识别
环境、
信道
语音增强、分离、
去混响
“形简意丰”

语音研究是典型的交叉学科

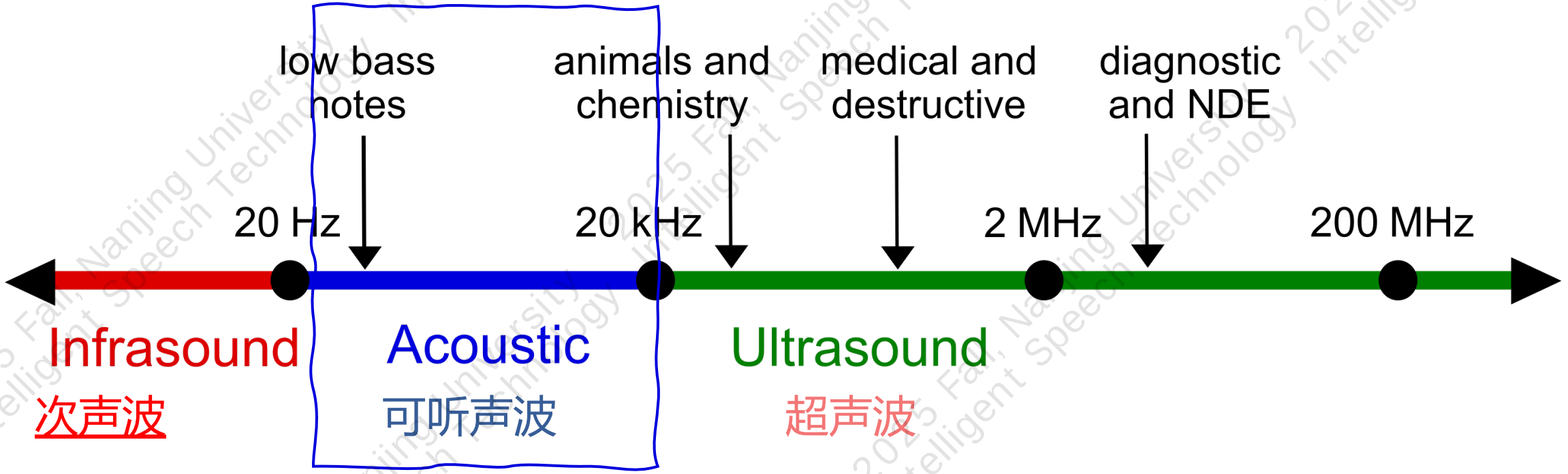


- **语言学 (Linguistics)**
  - 研究语言的形式、意义与语境
  - 探讨语言与社会、文化、历史等因素的关系
- **语音学 (Phonetics)**
  - 研究语音的物理、生物、心理特性
  - 关注声音本身，与意义无关
- **语音科学 (Science)**
  - 语音生成与发声机制研究
- **语音工程及应用 (Engineering&Applications)**
  - 语音编码
  - 增强与分离
  - 语音识别
  - 语音合成
  - 声纹识别

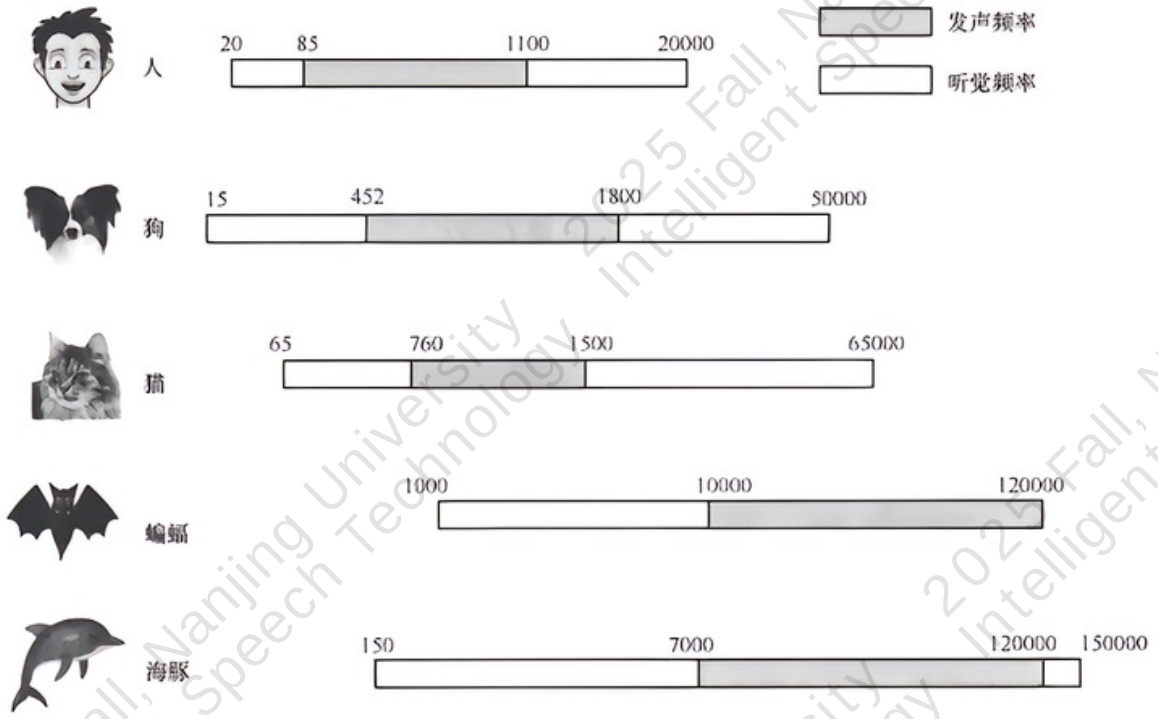
声音信号

声音是振动产生的声波，通过介质（气体、固体、液体）传播并能被人或动物听觉器官所感知的波动现象。

声波则可根据振动频率范围划分为次声波、可听声波 以及超声波



声音信号



次声波：< 20 Hz

- 来源：地震、火山爆发、核爆炸等
- 特性：波长较长，传播远、可绕过障碍物
- 应用：灾害预警、动物研究等

超声波：> 20,000 Hz

- 来源：蝙蝠、海豚、超声设备等
- 特性：波长短，能量集中，方向性好
- 应用：医学成像、工业检测、清洗

本次课程主要关注人类可感知的频率范围内的声波信号

声音信号

Sound: 声

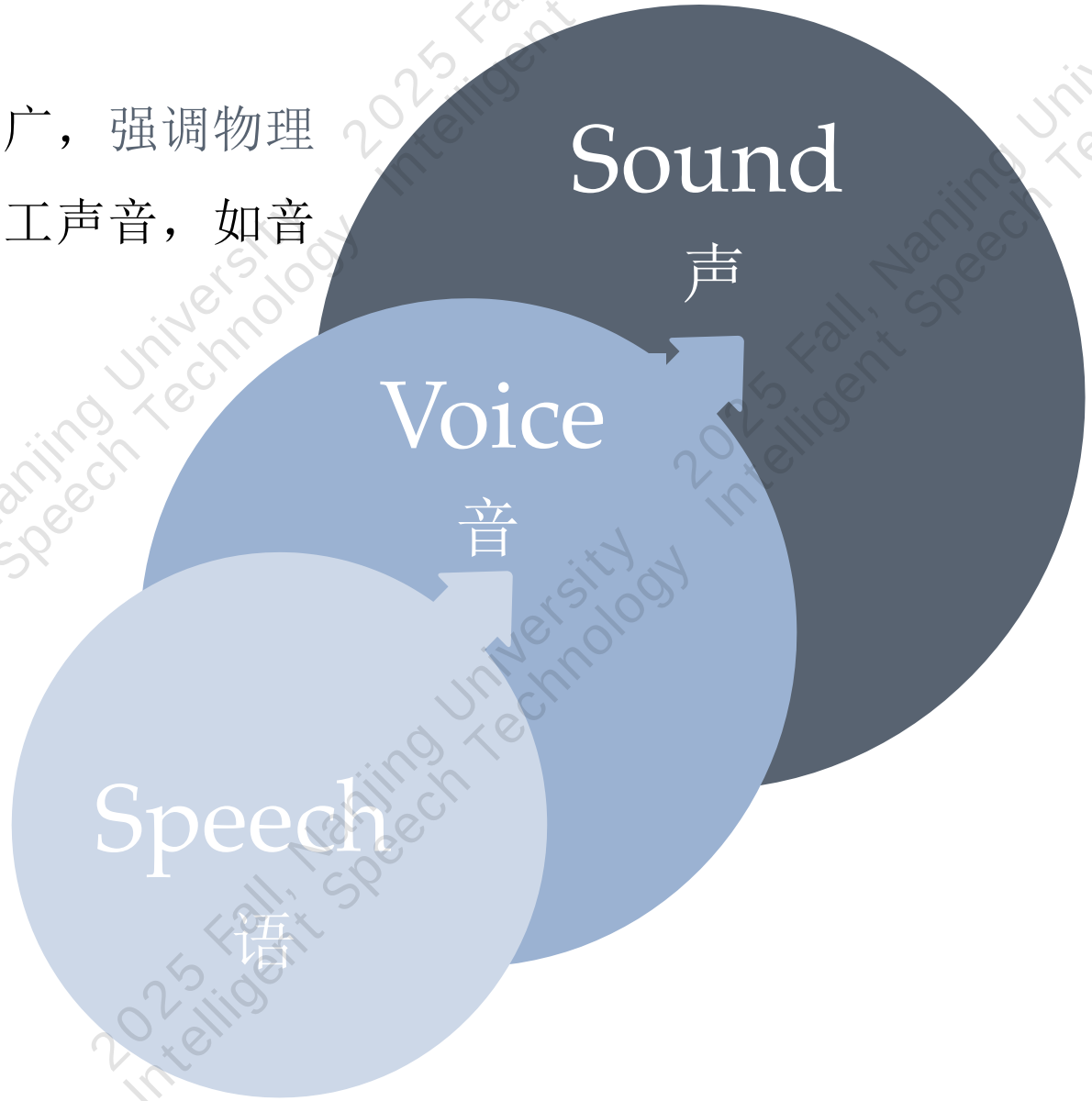
任何通过介质传播的振动，涵盖范围最广，强调物理现象本身；除人声外，还包括自然和人工声音，如音乐、鸟鸣、环境音

Voice: 音

由生物，特别是人类声带发出的声音，强调发声来源和个体属性

Speech: 语

人类为了沟通而组织发出的、具有意义的声音序列，是语言的载体



什么是语音信号？

语音信号是声音信号的一个种类，特指人发出的用来**传递信息**的声音信号，它是人在言语交流过程中由发音器官产生和听觉系统感知的**语言载体**。

你说的每一句话，每一个词，每一个字，都是语音信号

人发出的用来**传递信息**的方式有很多

手势

文字

• • •

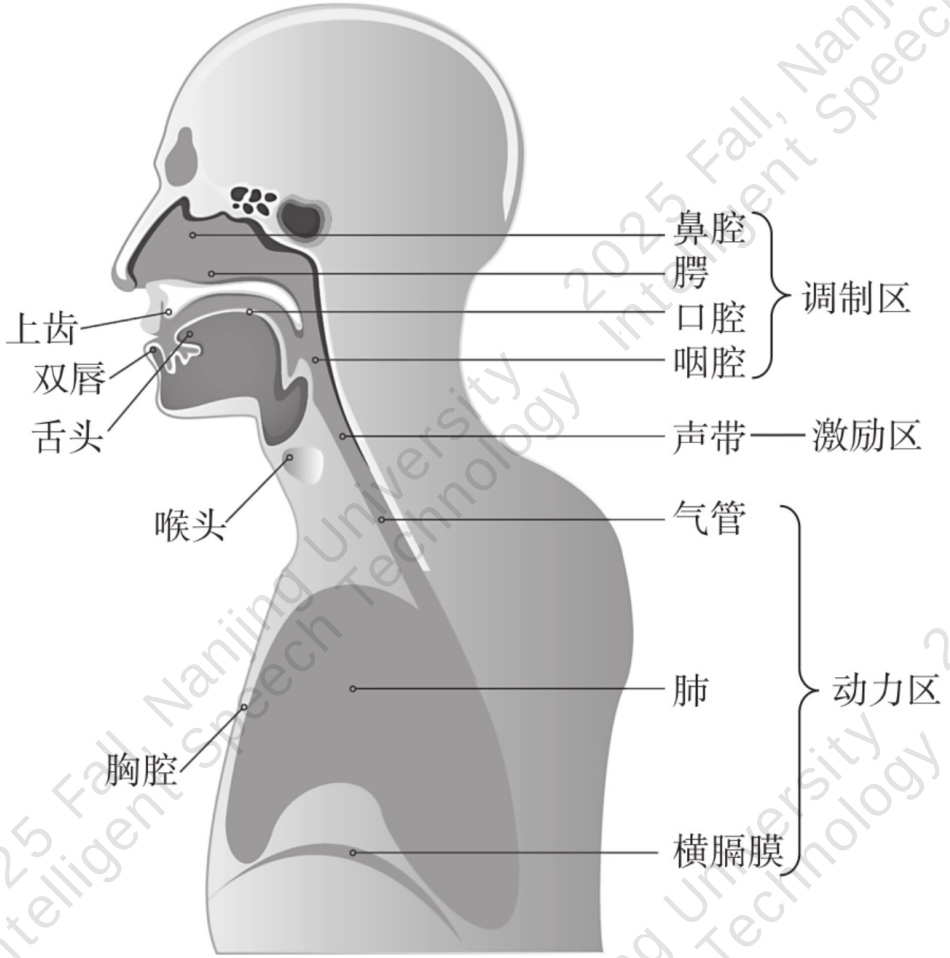
语音有哪些优势呢？

高效便捷	自然直观	特定用户友好	强互动性
➢快速传达信息，无需逐字输入。 ➢可在进行其他活动时使用，实现多任务处理。	➢接近日常交流方式，传达语气、情感更加丰富生动。 ➢复杂信息通过语音解释更易理解。	➢对视力障碍者友好，方便信息获取。 ➢低识字水平人群可降低信息获取门槛。	➢实时反馈，促进沟通理解。 ➢传递情感，增强人与人之间的情感连接。

语音是最自然、便捷的交互方式

人类的发音原理

人类的发音是一个复杂而精妙的过程，主要涉及动力区、激励区和调制区三个部分的器官协同作用。



调制区由舌头、上齿、双唇、咽腔等组成。

- 舌头的前后位置、高低位置以及卷曲程度的不同，可以产生不同的元音和辅音。
- 上齿和双唇可以通过开合、紧闭、突出等动作，与气流相互作用，产生不同的爆破音、摩擦音和鼻音等辅音。

激励区包括喉头、气管和声带。

- 喉头又称喉结，由软骨、肌肉和韧带组成，内部有声带
- 气流从肺部经气管上升至喉头，声带根据发音需要振动或不振动

动力区主要由胸腔、肺和横膈膜组成。

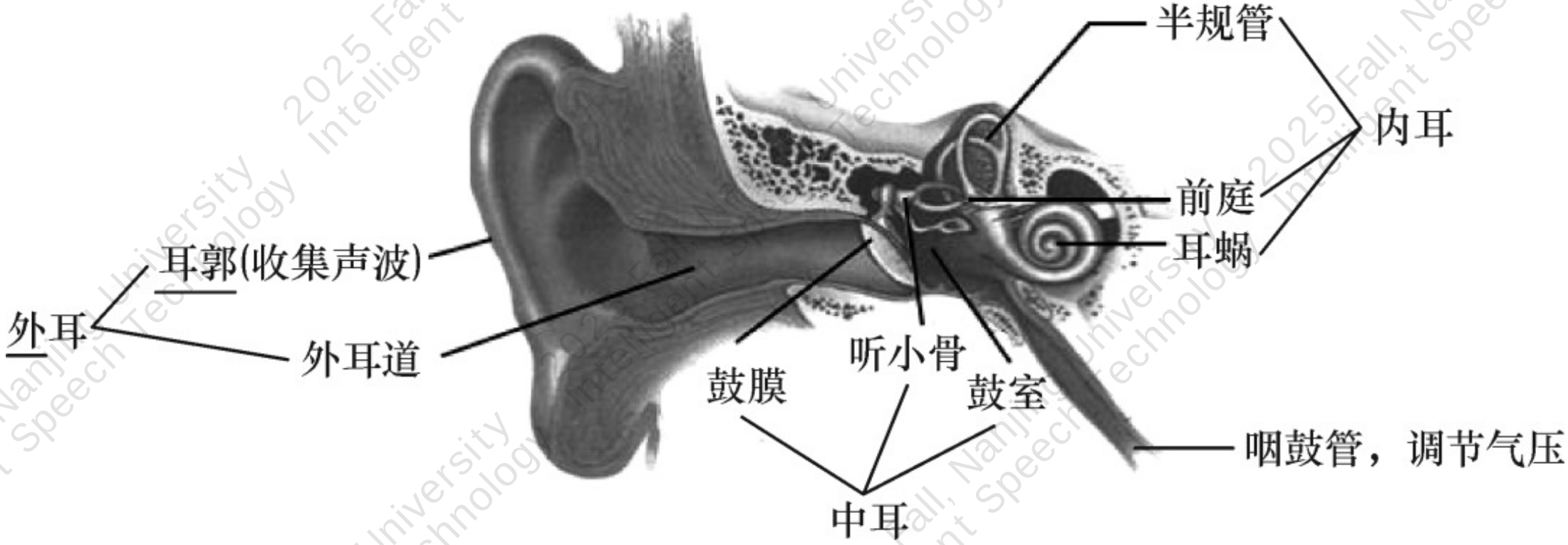
- 肺是呼吸主要器官，发音时肺部呼出气流提供动力
- 横膈膜位于胸腔和腹腔之间，在发音过程中，通过收缩和放松来调节气流的强弱和稳定性。

人类的发音原理

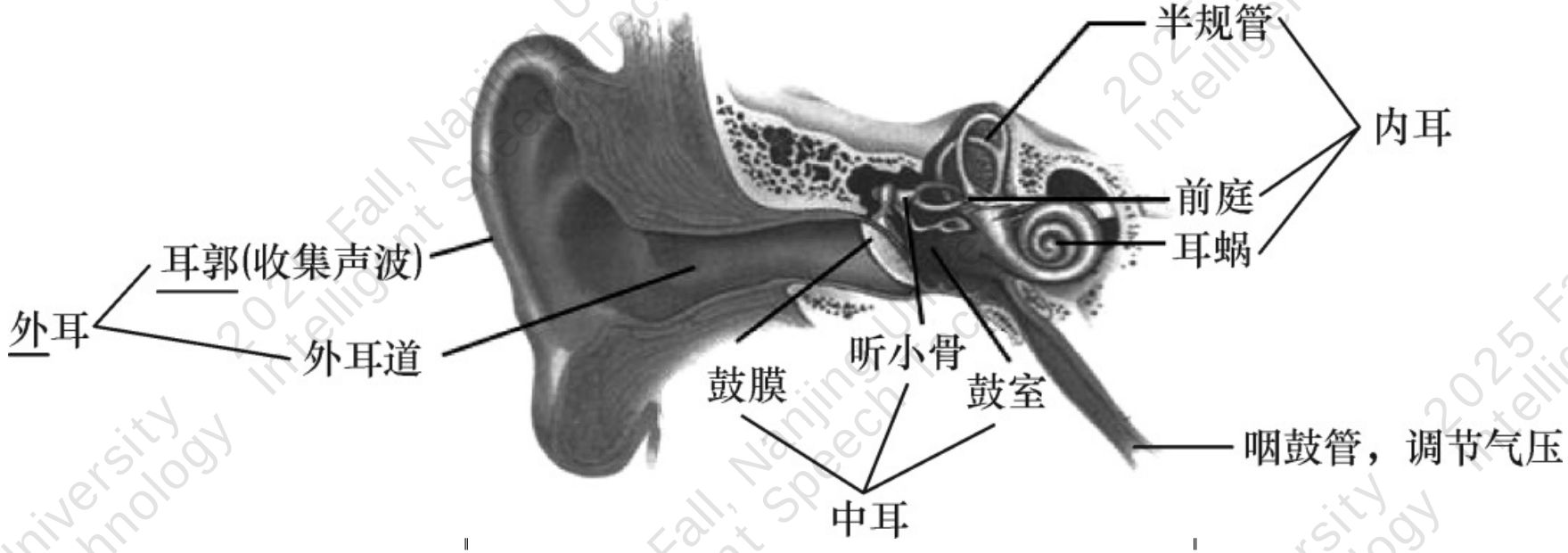
来源：https://www.youtube.com/watch?v=JF8rlKuSoFM美国国立卫生研究院

人类的听觉机制

人耳由外耳，中耳，内耳构成



人类的听觉机制



收集声音，耳廓收集周围声音并确定来源方向。

保护作用，外耳道弯曲结构和耳毛阻止异物进入。

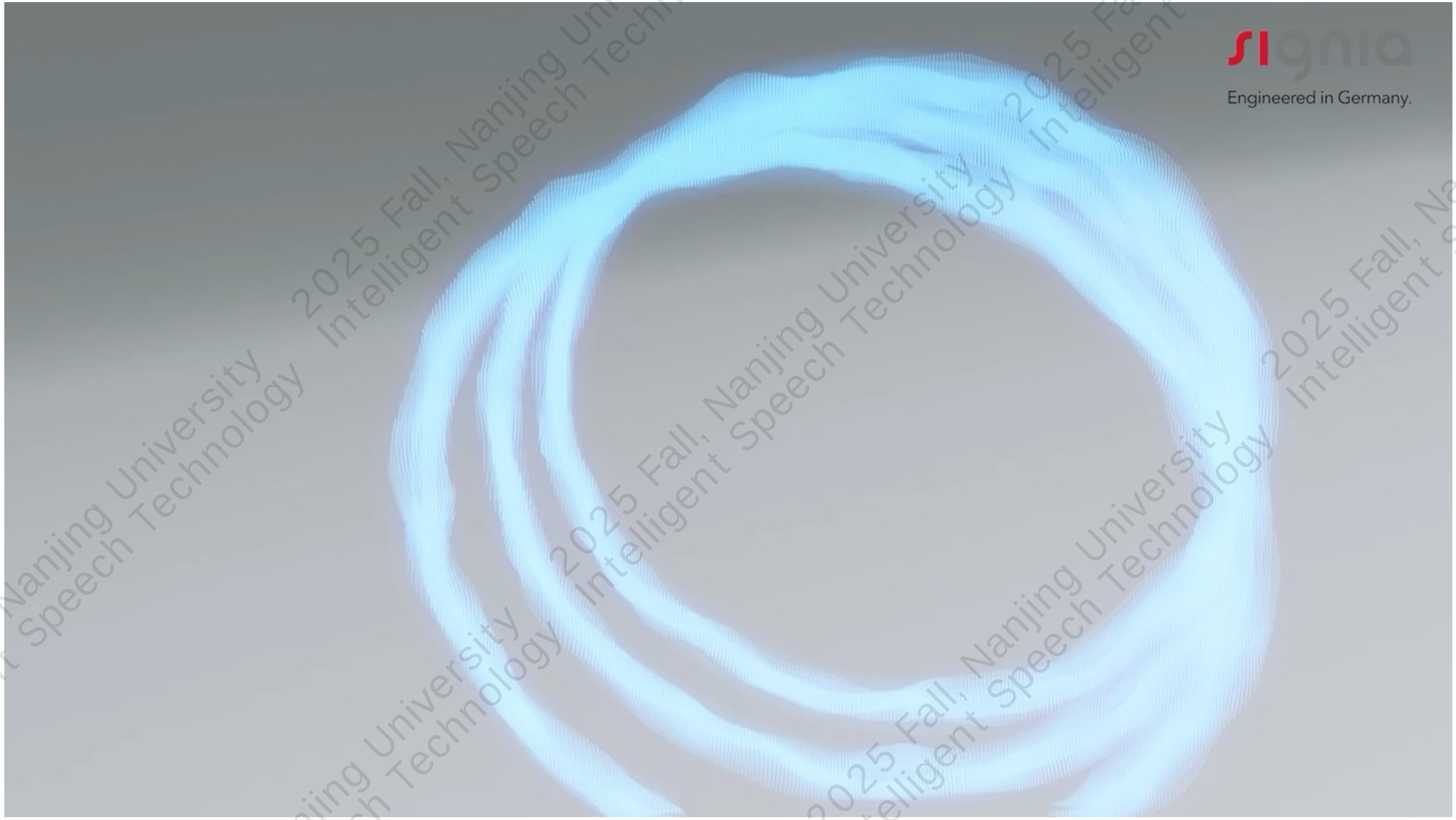
传导声音，鼓膜和听小骨将声音传递到内耳，放大声音能量。

平衡压力，咽鼓管调节中耳气压与外界平衡。

感知声音，耳蜗将声音振动转换为神经信号。

维持平衡，半规管和前庭感知头部运动和位置变化，维持身体平衡。

人类的听觉机制



来源： https://www.bilibili.com/video/BV1Cg4y1z7mn 西嘉助听器

智能语音技术的发展史

说话机器：肯佩伦，1796年

奥匈帝国发明家沃尔夫冈·冯·肯佩伦的说话机器

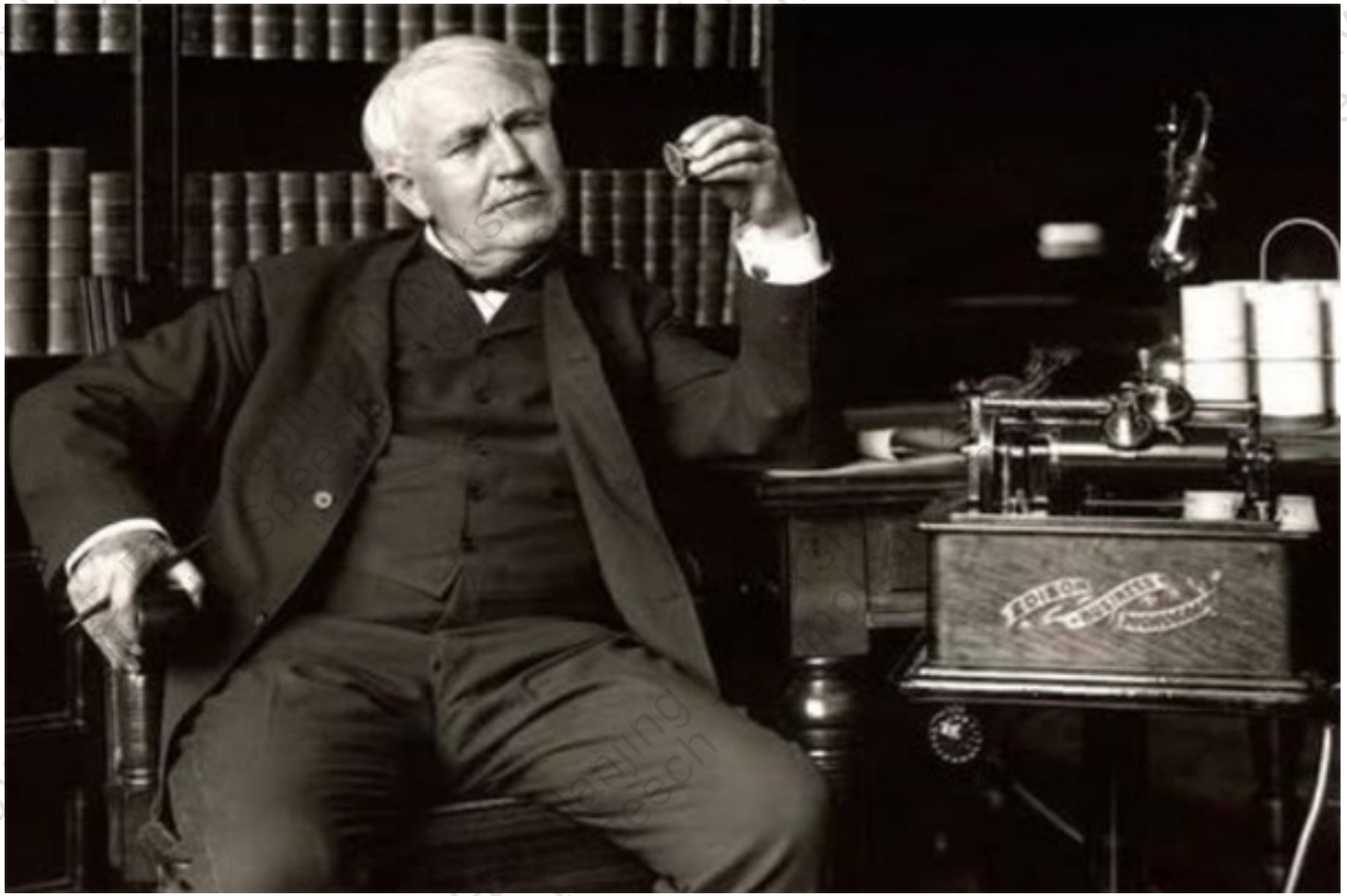
仿照人类发音机制

厨房风箱	为肺提供气流
单簧管吹口	作为嘴
风笛簧片	作为声门

肯佩伦的说话机器的一个复制品，建造于2007–2009年，在德国萨尔布吕肯萨尔州大学语音学系。

智能语音技术的发展史

留声机：爱迪生，1879年



物理原理：

声音的本质是振动，只要记录振动的特性，并加以模拟，就可以发出同样的声音

留声机的发明为声音的记录、保存和传播带来了革命性的变化，也为后来录音技术的发展奠定了基础。

智能语音技术的发展史

电话的发明：贝尔，1876年

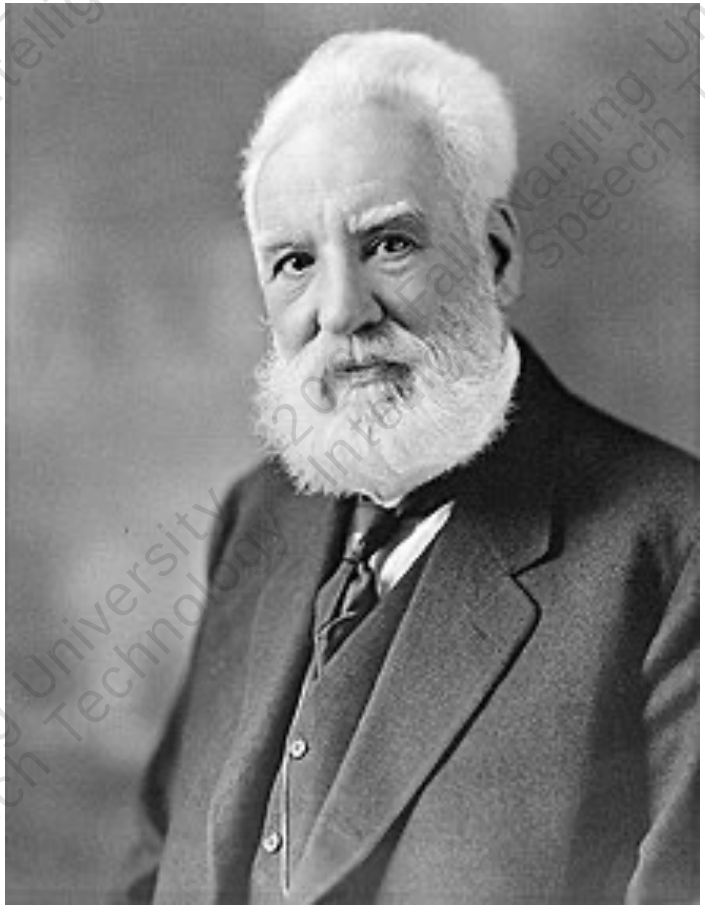
贝尔一声都在于声学、语音学打交道，除了发明家的身份外，他还是声学生理学家和聋哑人语教师。他的母亲和妻子都是聋人。

发明：电话、改良留声机、听力测量设备等等

声音强度的一个重要衡量指标： 分贝（dB），便是为了纪念贝尔而命名的

贝尔在1885年的一次实验中把自己说话的声音录在一张纸音盘上，成为人类史上的第一张用声音签名的音盘。

亚历山大·格雷厄姆·贝尔



来自1885 年的贝尔的录音

智能语音技术的发展史

初代语音识别系统：Audrey，1952年

Audrey 系统是最早的基于电子计算机的语音识别系统。由 AT&T 贝尔实验室于 1952 年开发，可以识别 10 个英文数字（“0” 到 “9”）

该系统只能识别有限的数字，且对使用者有一定要求：需要使用者进行一些调整以适应系统，对熟人的准确度高达 90％以上

标志着自动语音识别领域的开端

智能语音技术的发展史

初代语音合成系统：Voder，1939年

贝尔实验室的 Voder（Voice Operation Demonstrator 的缩写）是世界上第一台电子语音合成器，并由荷马·达德利（Homer Dudley）改进。这台机器能够说出“下午好，广播听众”等语句。

Voder 系统在当时并未被广泛采用，但它为后续的语音合成技术的发展奠定了基础，推动了该领域的进步。

1939年纽约世界博览会展出的键盘式合成器

智能语音技术的发展史

初代商用语音合成系统：DECtalk，1984年

DECtalk 是一种数字语音合成技术，由美国 Digital Equipment Corporation（DEC）公司开发, 迅速成为商用市场上首批高质量语音合成器之一。

在 20 世纪 80 年代后期，霍金开始使用 DECtalk 系统。

DECtalk lets micros read messages over phones

BY PEGGY ZIENTARA
Senior Editor

The use of the telephone as a "universal terminal" came one step closer to reality when Digital Equipment Corporation (DEC) introduced a voice synthesizer unit that can "read" aloud ASCII text messages.

DECtalk, which the company says will be available in March at a price of $4000, allows you to call in to a personal computer from anywhere in the world by touch-tone telephone (including pay phones), to hear electronic mail messages stored in memory.

DECtalk can also "tell" you what you just typed into memory on the computer keyboard if you simply press a punctuation mark key and the Return key.

Business applications are expected to include access to company data bases by traveling sales representatives; access by lawyers to public and private computer files during trial breaks; direct customer transactions with banks and stock brokerage firms; direct catalog and retail sales; and access to transportation schedules, according to Ed Kramer, vice-president of corporate marketing for DEC.

Other applications will include aid for the speech-impaired (now in place at

via any touch-tone phone.

The DECtalk hardware, which fits easily under a 12-inch monitor, will feature eight different types of voices, including that of an old man, a deep-voiced man, a typical middle-aged man, two different women and a child. The unit attaches to a personal computer via an RS-232C serial interface.

DEC claims to have licked the traditional voice technology trade-offs of having either good voices with limited vocabularies or unlimited vocabularies with poor voices. DEC's unit has a "virtually unlimited vocabulary," Kramer said, and the speaking rate can be adjusted from 120 to 350 words per minute. Tone and modulation can also be regulated.

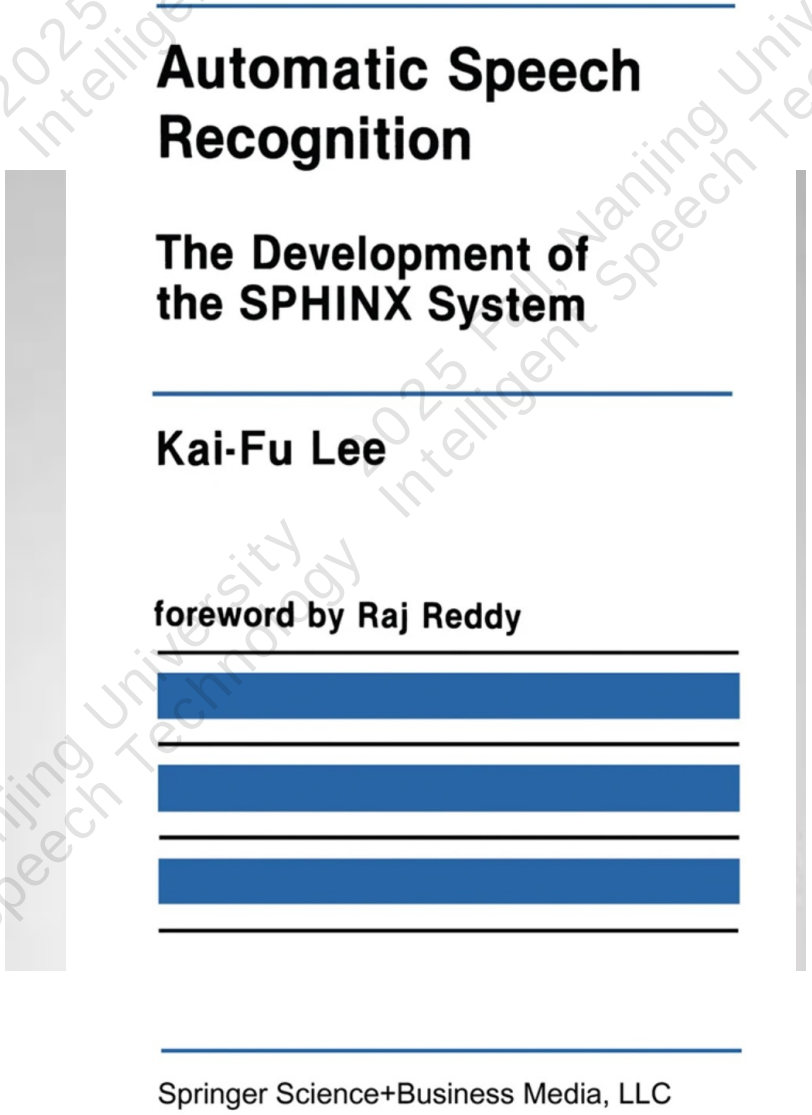
Heuristics — interpreting word context to improve pronunciation — and a computer model of the human vocal tract allow DECtalk to produce high-quality voices, in contrast with traditional stored digitized or synthesized voice technology, Kramer explained.

The unit accepts ASCII text, which it applies to an internal dictionary and a set of letter-to-sound rules indicating pronunciation and context interpretation to

1984年美国InfoWorld报道DECTalk

上世纪90年代的语音助手

https://www.youtube.com/watch?v=PJ_KCTsOCrs



目录

CONTENTS

智能科学与技术学院
School of Intelligence Science and Technology

语音交互链路（Speech Chain）

听

接收信息输入

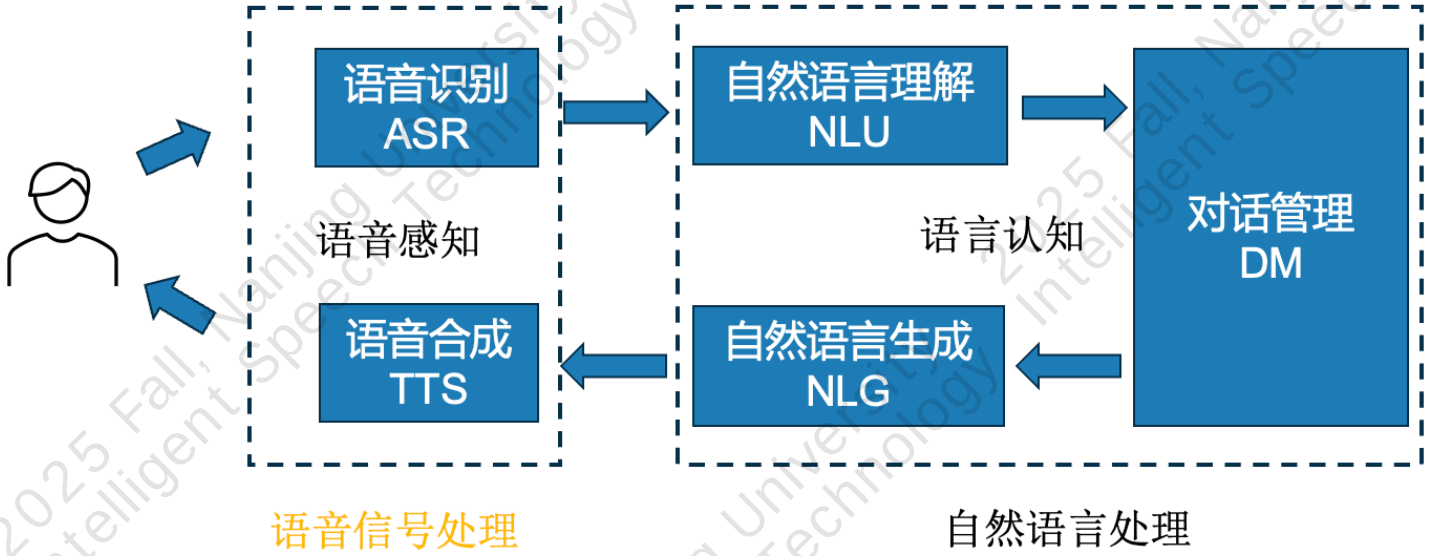
人类交流

说

产生信息输出

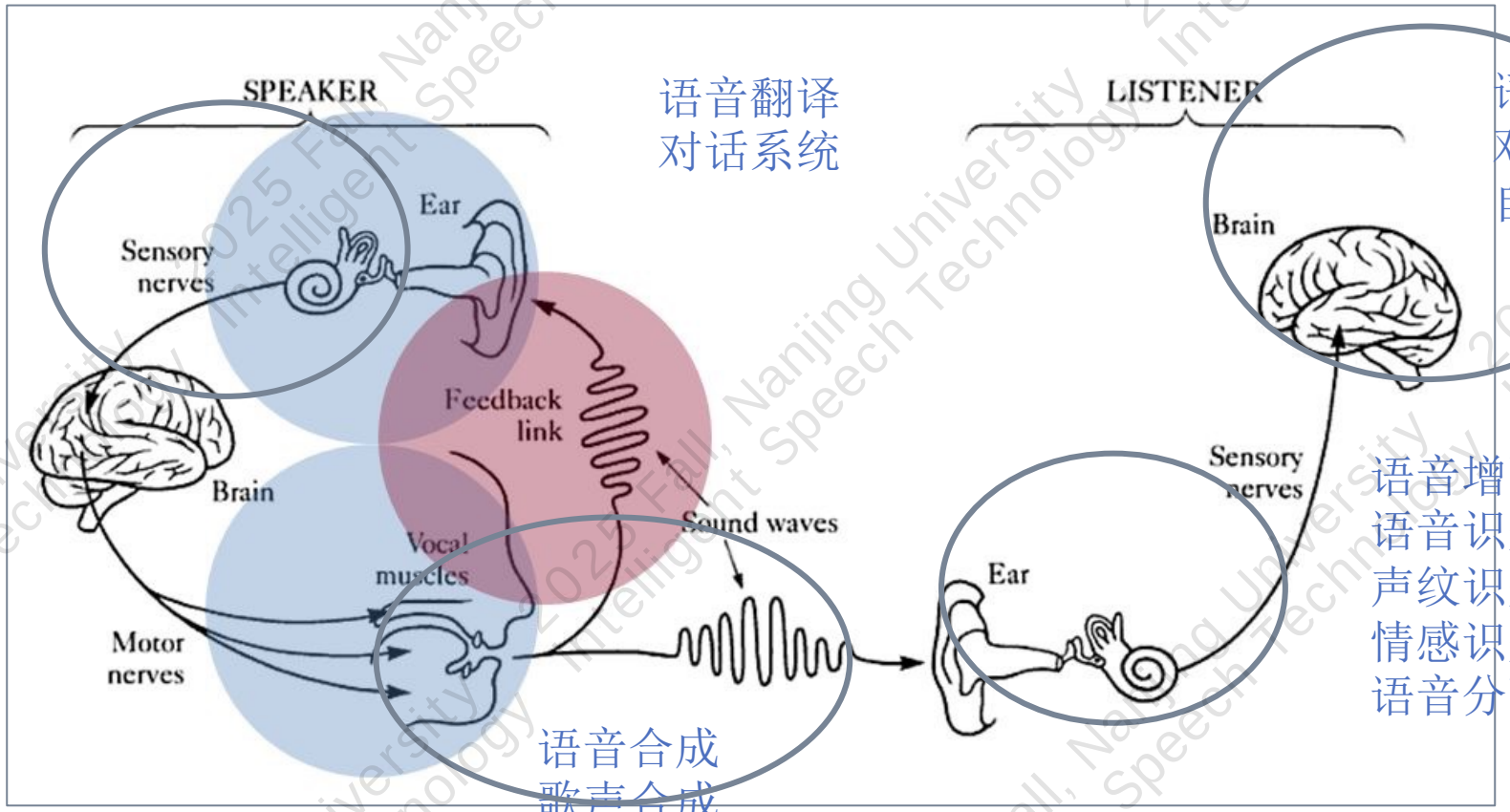
理解

信息整理和加工



语音交互链路（Speech Chain）

神经科学
听觉物理学



语音翻译
对话系统

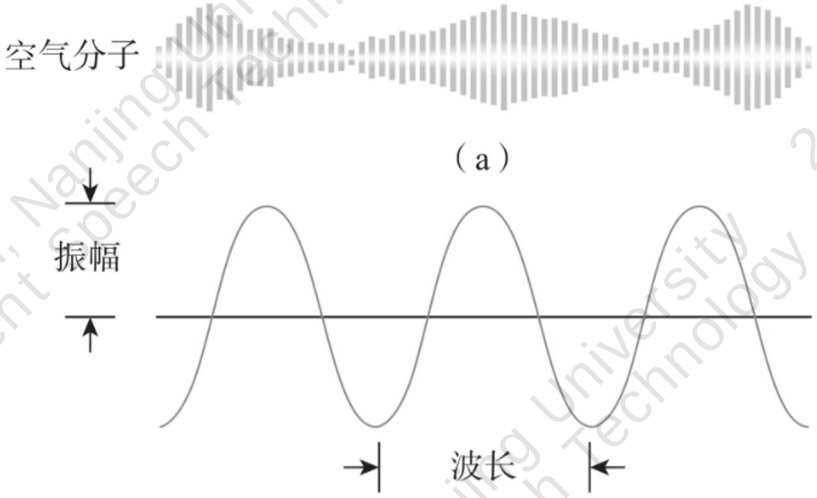
语义理解
对话管理
自然语言生成

语音增强
语音识别
声纹识别
情感识别
语音分离

语音合成
歌声合成

语音信号的表示形式

声带振动产生声波，声波在空气中传播，引起空气分子的振动，从而形成了空气疏密相间的周期性变化

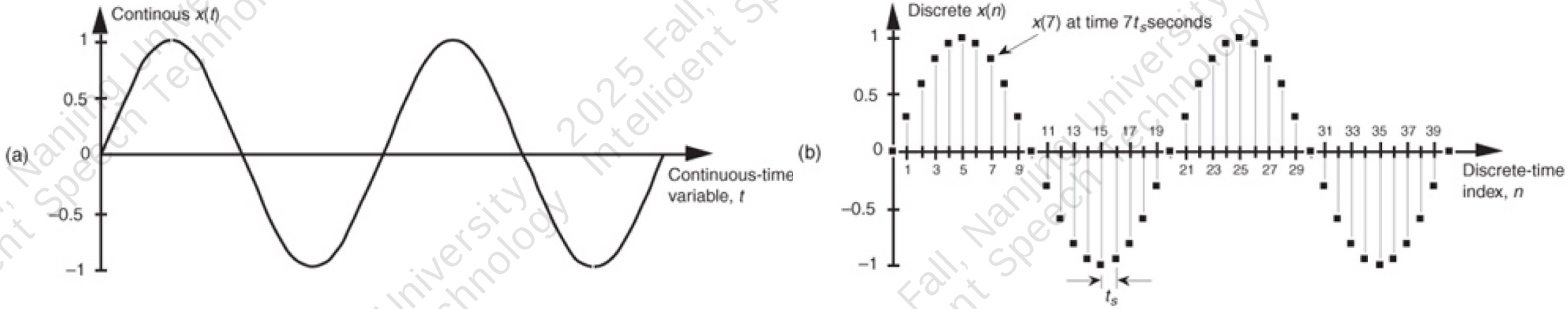


每隔一个非常短的时间（如1/16000秒）记录一次该点处的语音信号，这一过程称为采样，每个记录值称为一个采样点，一秒钟内的采样次数称为采样频率

将模拟信号转换成计算机可处理的数字信号

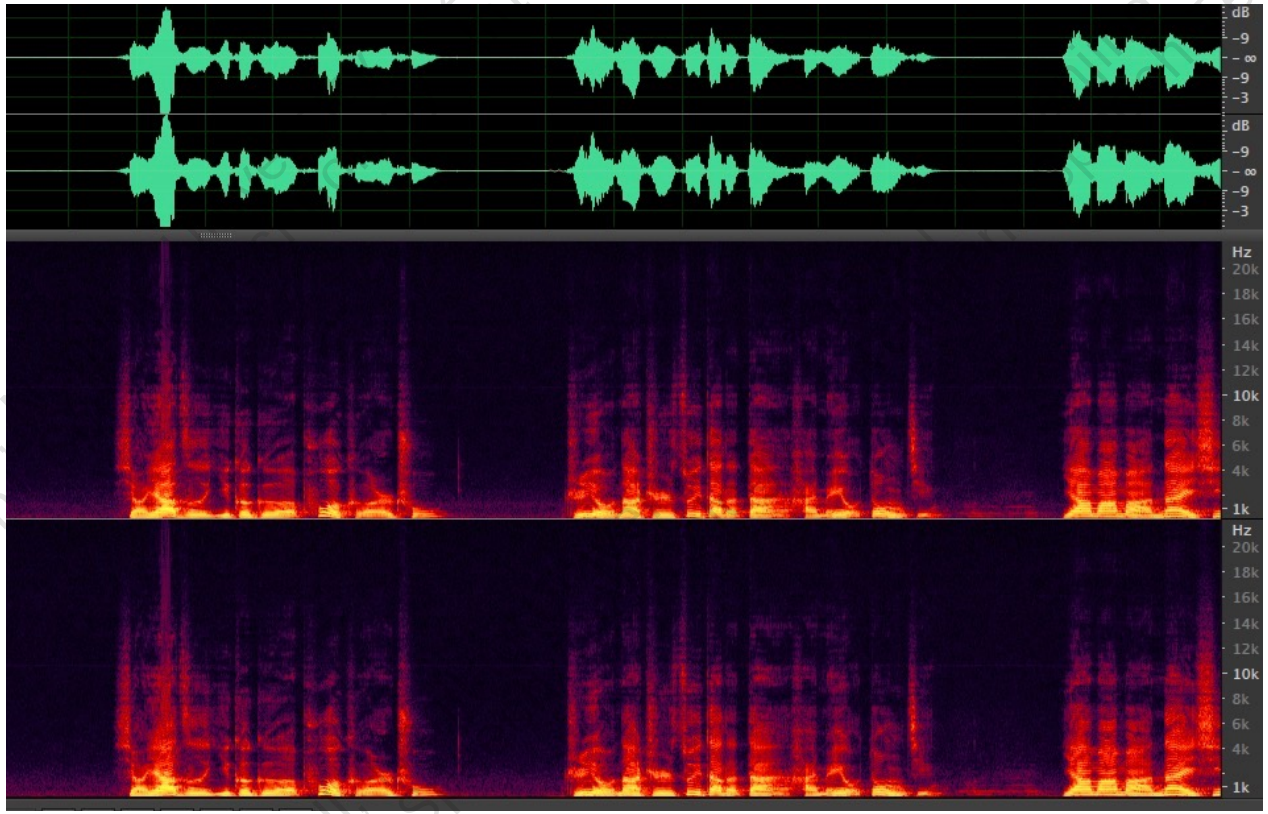
模拟信号到数字信号转化（analog-to-digital converter，ADC）

在科学和工程中，遇到的大多数信号都是连续的模拟信号，例如电压随着时间的变化，一天中温度的变化等等，而计算机只能处理离散的信号，因此，必须对这些连续的模拟信号进行转化，通过采样和量化，转换成数字信号。

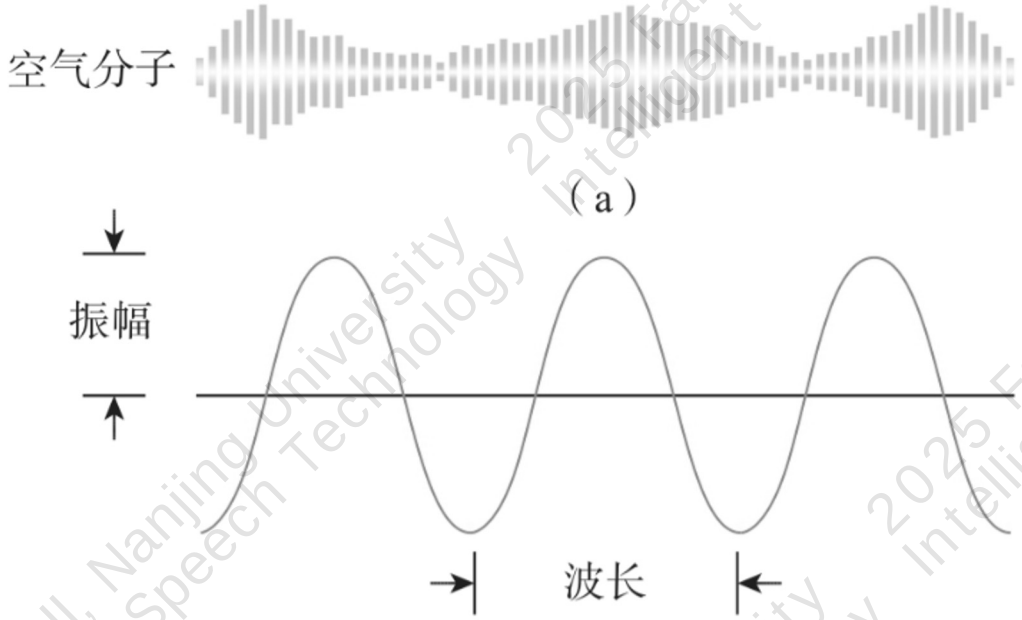


语音信号的表示形式

- 采样率、量化位数、通道数
  - 16KHz, 16bit, Mono
  - 8KHz, 8bit, Mono
  - 44.1KHz, 16bit, stereo
  - 48KHz,24bit, stereo
  - …
- 数据率（速率/码率）
  - 数据率=采样频率×量化精度×声道数
  - 5分钟48KHz,24bit, stereo的音频数据率是多少？



语音信号的表示形式

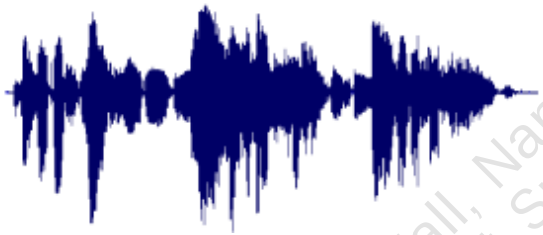


振幅即波形的最高点（或最低点）与基线间的距离，它表示了声音音量的大小。

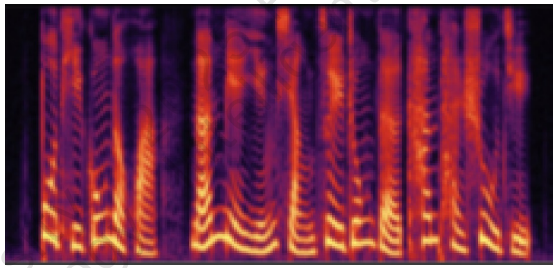
周期是波形中两个相邻波峰之间的距离，它表示完成一次振动过程所需的时间， 其大小体现了振动的速度。

频率是周期的倒数，周期越短，频率越高。频率的单位为赫兹(Hz)。

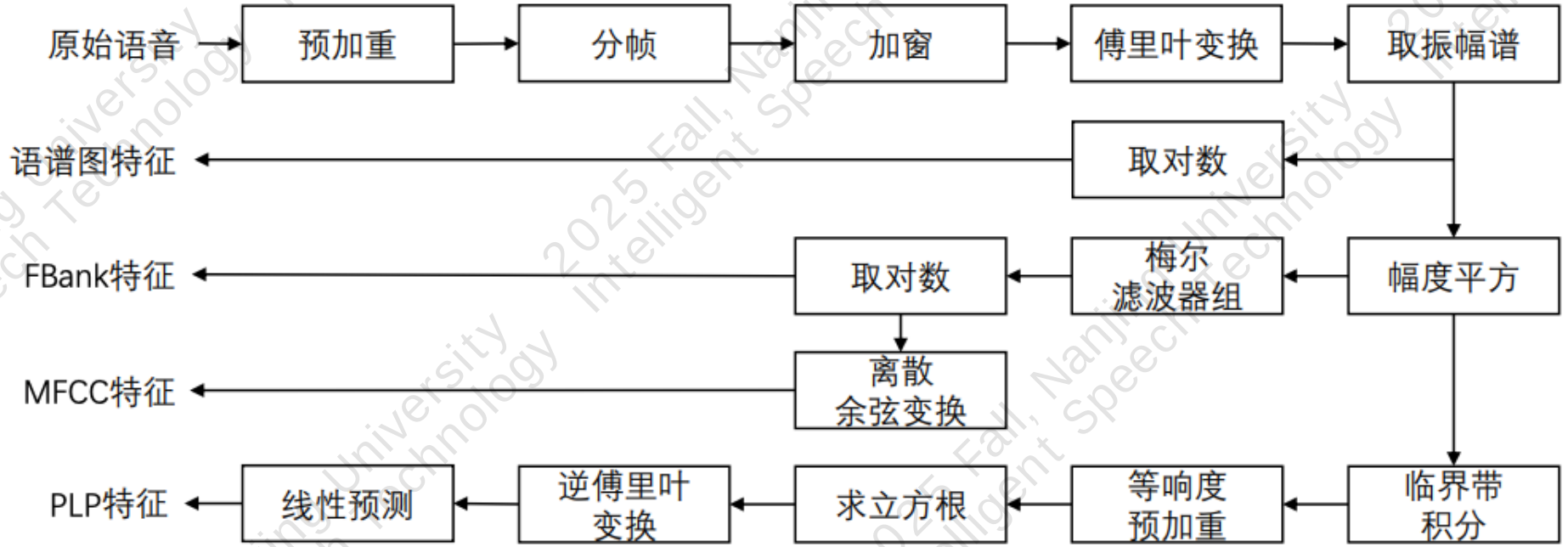
语音信号的表示形式



语音信号时域



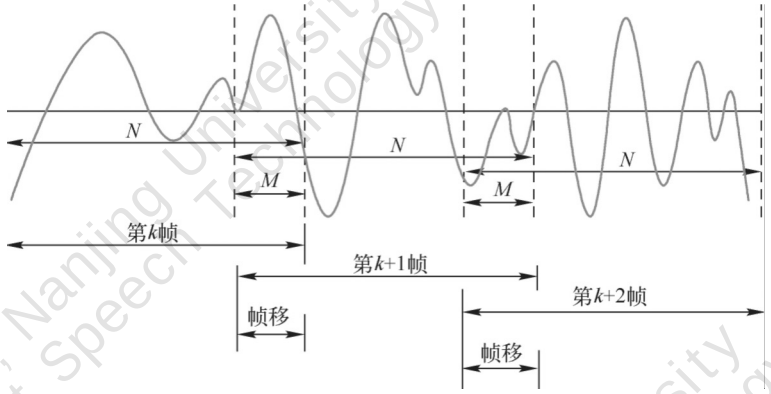
频谱特征频域



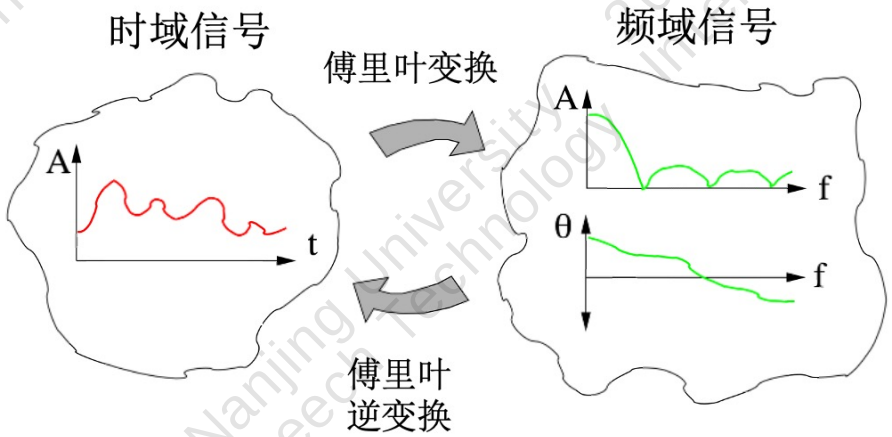
语音信号的表示形式

这里简要介绍其中两个重要步骤

分帧：语音是短时平稳信号，先分割成有重叠的帧，再进行后续处理



傅里叶变换：将一个在时域表示的函数转换为在频域表示的函数，从而揭示出信号所包含的频率成分及其强度

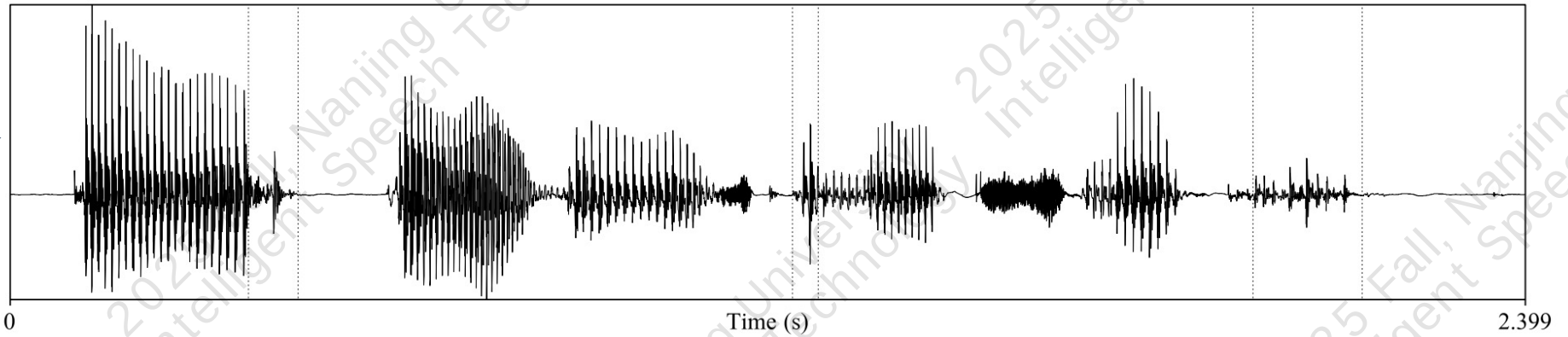


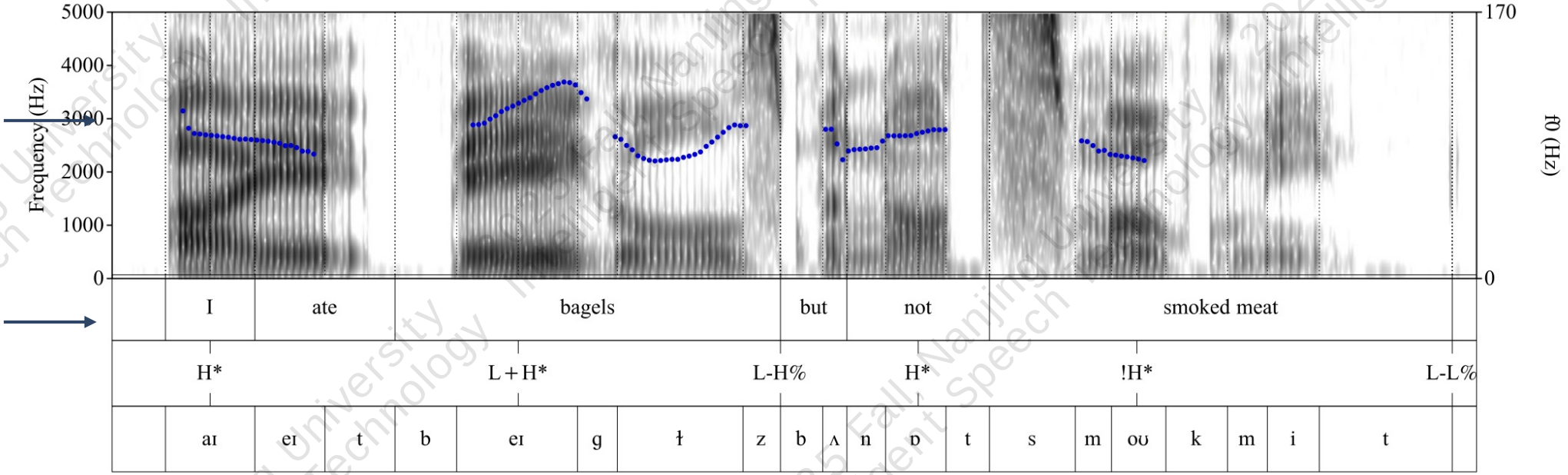
语音信号的表示形式

语音波形 →

语音频谱 →

说话内容 →





语音编码技术

目标：压缩语音信号，助力语音的数字化存储与传输

出发点：

- 更好的语音质量
- 更强的抗干扰性，并易于进行加密
- 节省带宽，能够更有效地利用网络资源
- 更加易于存储和处理

可行性：

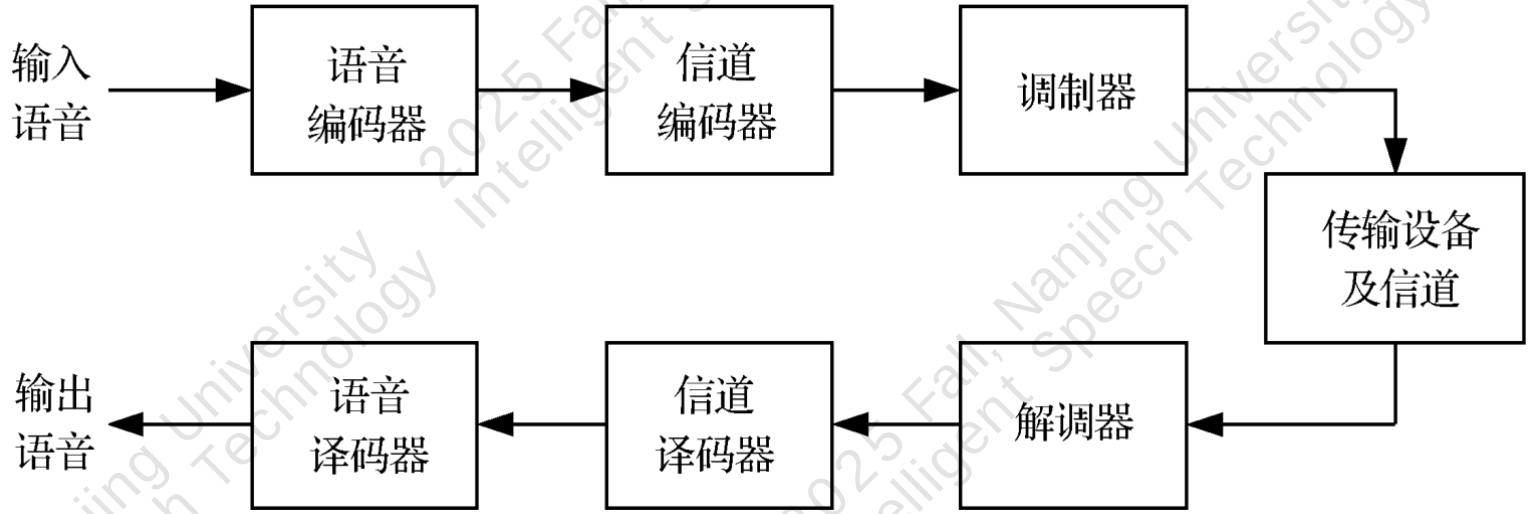
- 声音信号中包含有大量的冗余信息
  - 邻近样本之间有很大的相关性
  - 长时（几十秒）自相关性
  - 话音间歇（静音）
- 可以利用人的听觉感知特性进行压缩

语音编码技术

语音编码系统应用

数字语音录放系统：编码，存储，回放

数字电话通信系统：编码，传输，译码



语音编码技术

传统语音编码分类

波形编码：

基于对语音信号波形的数字化处理，试图使处理后重建的语音信号波形与原语音信号波形保持一致

优点　实现简单、语音质量好、适应性强，有成熟的技术实现方法；

缺点　压缩程度不高、增加压缩比例严重影响质量

参数编码：

通过构造发声模型作为基础，用一套模拟声带频谱特性的滤波器系数和若干声源参数来描述这个模型

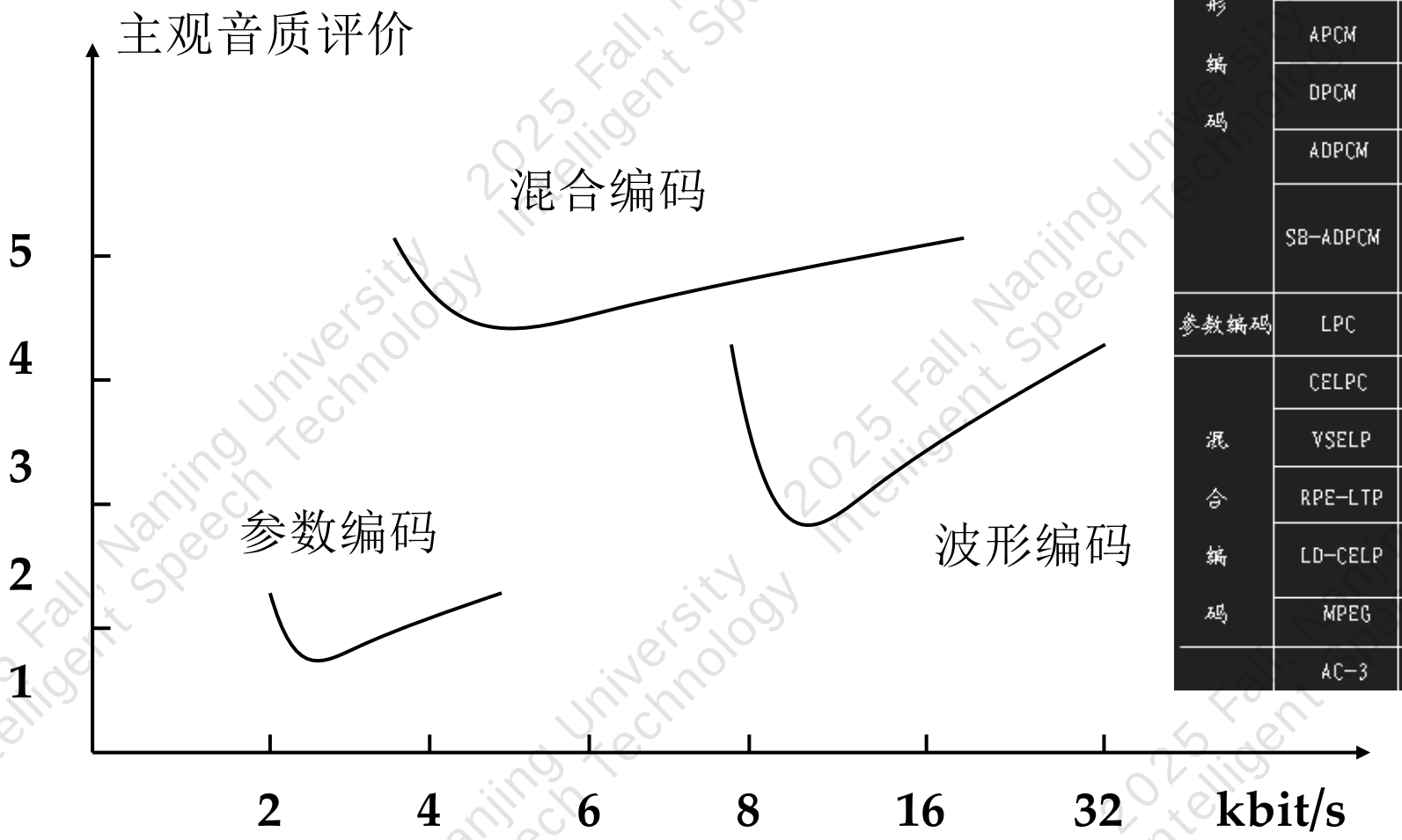
优点　语音编码速率较低（2～9.6kbit/s），压缩率高；

缺点　合成语音质量较差，实现的复杂度高

混合编码：

结合波形编码和参数编码，既包含若干语音特征参量，又包括部分波形编码信息。

三种压缩编码的性能比较



	算法	名称	数据率	标准	应用	质量
波形编码	PCM	均匀量化			公用网 ISDN 配音	4.0-4.5
	μ(A)	μ(A)	64kb/S	G.711		
	APCM	自适应量化				
	DPCM	差值量化				
	ADPCM	自适应差值量化	32kb/S	G.721		
	SB-ADPCM	子带-自适应差值量化	64kb/S	G.722		
			5.3kb/S 6.3kb/S	G.723		
参数编码	LPC	线性预测编码	2.4kb/S		保密话声	2.5-3.5
混合编码	CELPC	码激励LPC	4.8kb/S		移动通信	4.0-3.7
	VSELP	矢量和激励LPC	8kb/S		语音邮件	
	RPE-LTP	长时预测规则码激励	13.2kb/S		ISDN	
	LD-CELP	低延时码激励LPC	16kb/S	G.728 G.729		
	MPEG	多子带 感知编码	128kb/S		CD	5.0
	AC-3	感知编码			音响	5.0

技术发展史

模板匹配

动态时间规整算法：与样本库的语音进行时间轴对齐

统计模型

经典的GMM-HMM 算法基于传统统计模型（李开复示例）

深度学习

传统的统计模型采用神经网络替代，语音识别是深度学习大展神威的第一个领域

评价标准

ASR 系统质量常通过比较假设转录和参考转录衡量，最常用指标为词错误率（WER），S、I、D 分别代表替换、插入、删除的数量

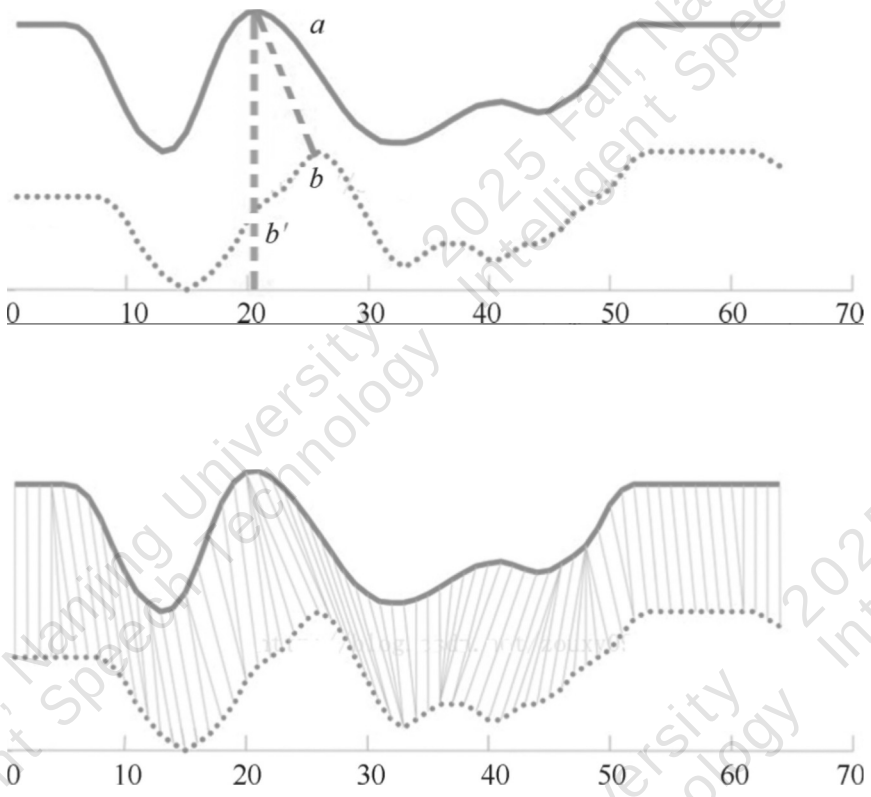
$$WER = \frac{S + I + D}{N} \times 100\%$$

标准答案：远大/的/理想/和/抱负
ASR结果： 远大/的/理想/和/报复/好

- 与参考转录相比，“抱负”被替换成了“报复”，即替换数 S = 1。
- 没有删除项，即删除数 D = 0。
- “好”是多余的插入项，即插入数 I = 1。

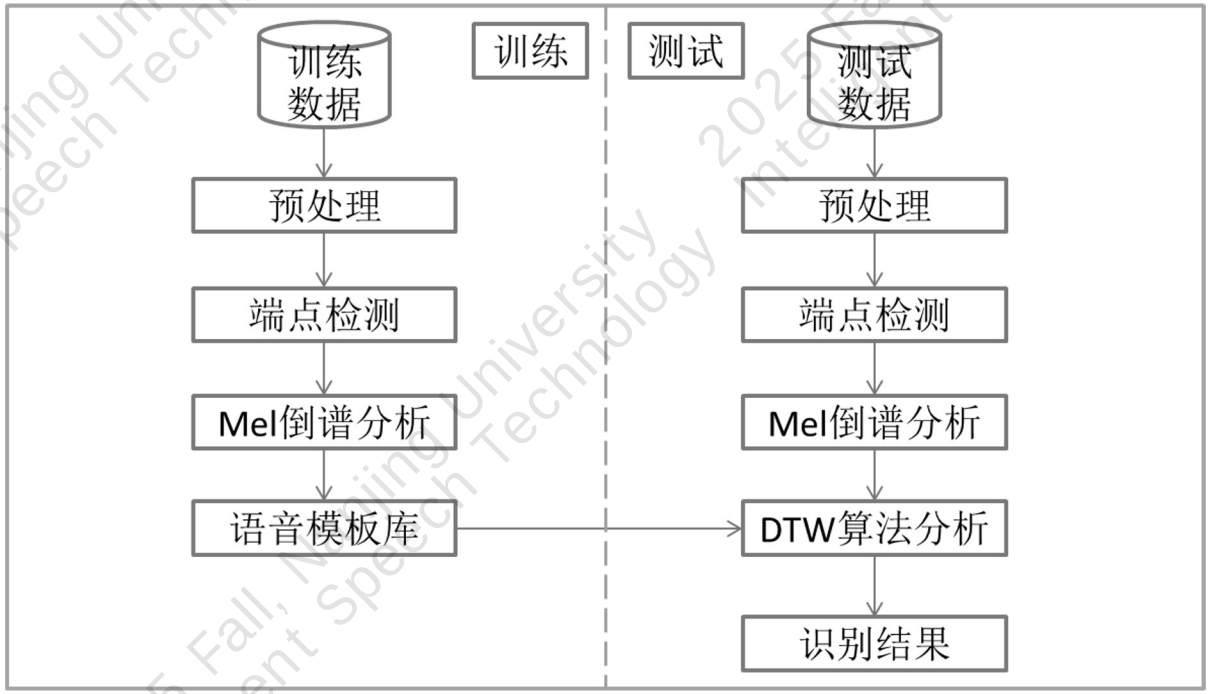
$$\text{WER} = \frac{1 + 1 + 0}{5} \times 100\% = 40\%$$

基于模板匹配的语音识别系统

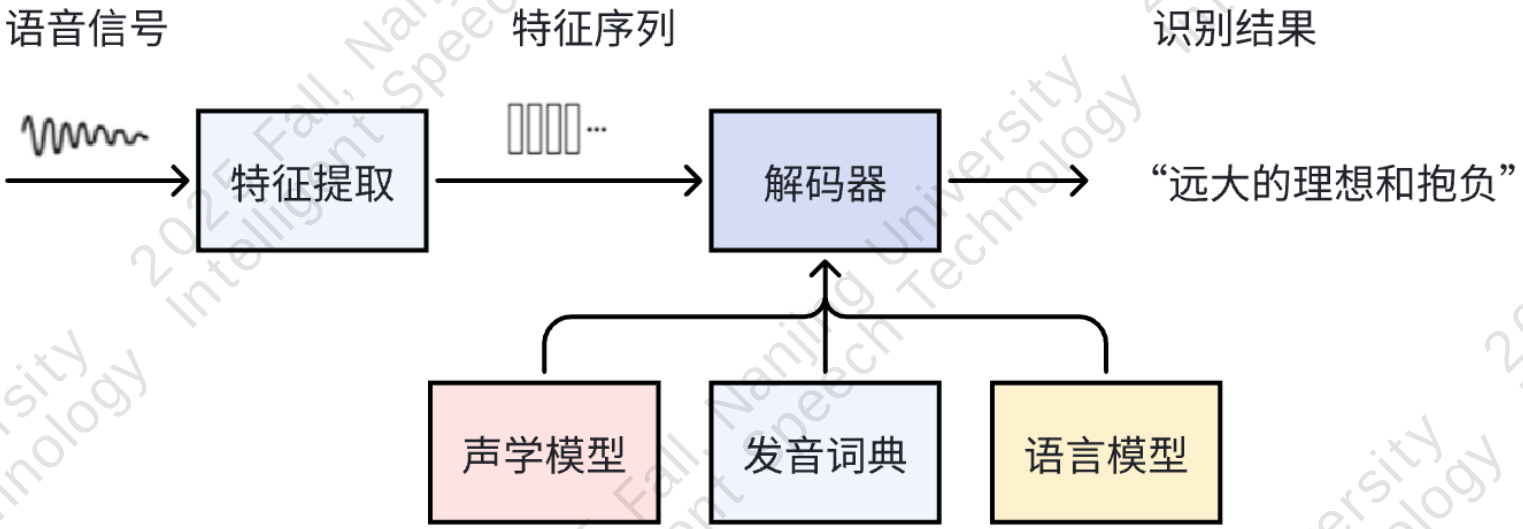


同一个单词“pen”的两个波形的DTW 示意图

动态时间规整(Dynamic Time Warping，DTW)，由日本学者Itakura提出，是一种衡量两个长度不同的时间序列的相似度的方法。



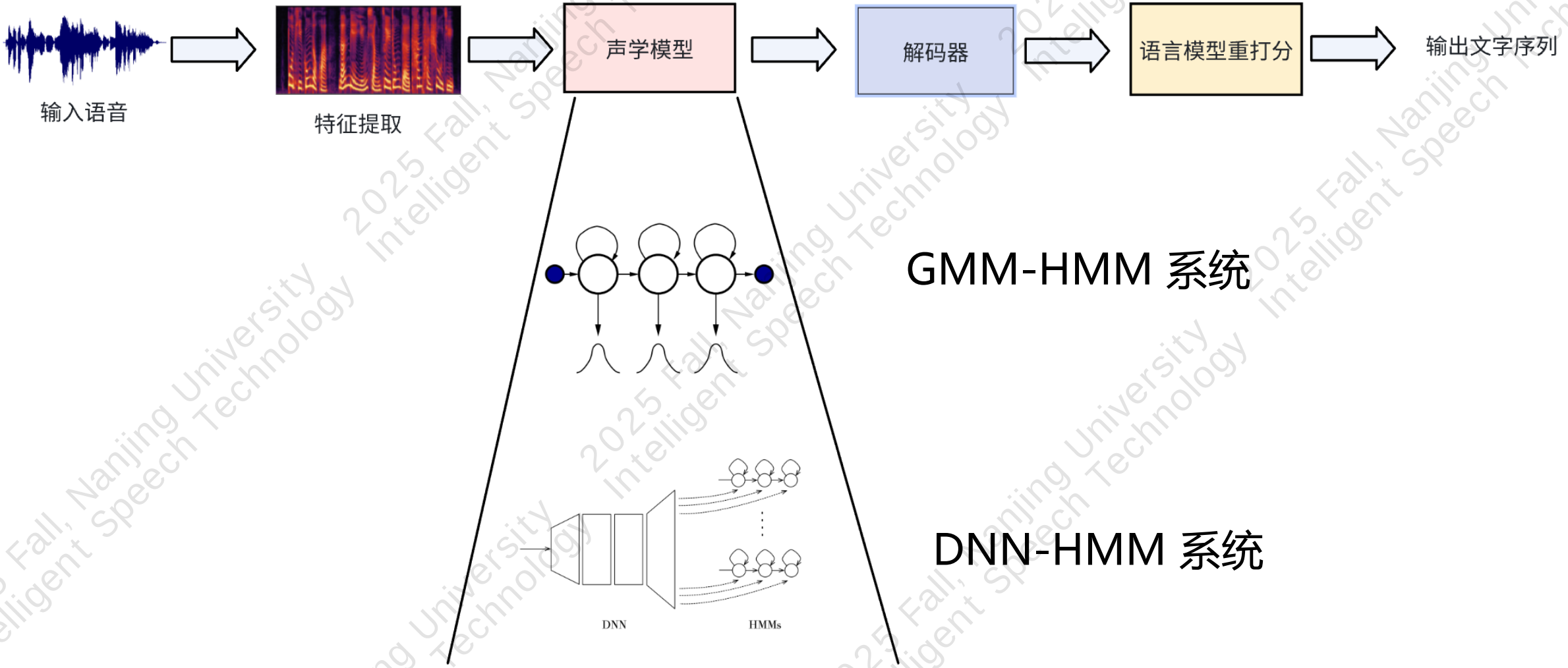
基于统计模型的语音识别系统



建模单元：音素 （中文为声母，韵母）

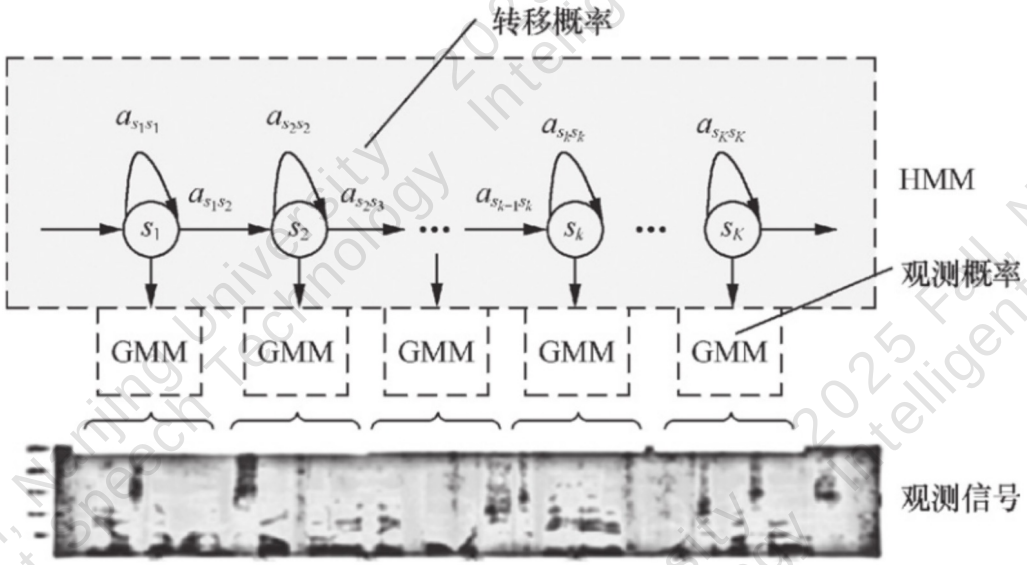
- 声学模型：特征序列 -> 音素序列 （ /yuan3-da4-de-li3-xiang3-he2-bao4-fu4/
- 发音词典：/yuan3 – da4/ -> “远大”
- 语言模型：远大的理想和 /bao4-fu4/ 抱负 报复
- 解码器：综合以上信息，搜索最优的输出序列

传统的语音识别系统



传统的语音识别系统

声学模型将声学特征映射到发音单元，为后续的语言模型提供可能的语音单元候选



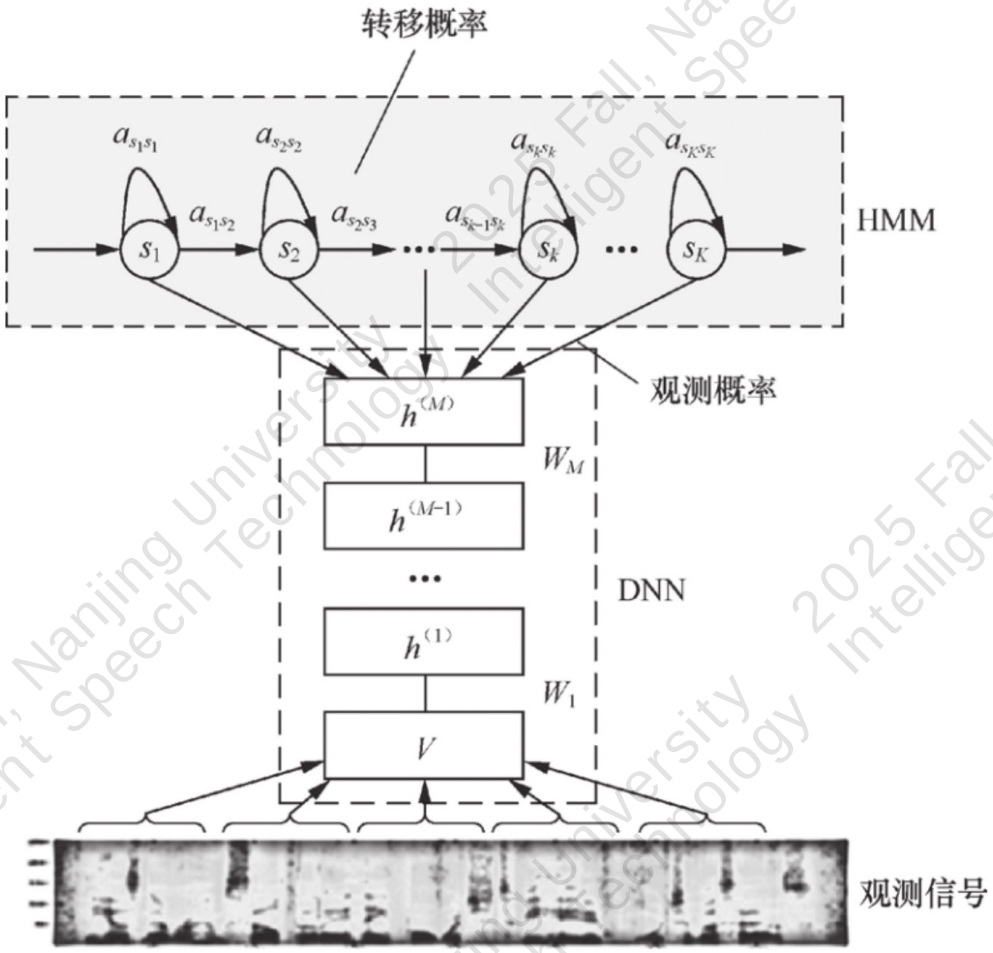
GMM-HMM 框架

HMM（隐马尔科夫模型）假设语音信号是由一系列隐藏的状态产生,一个发音单元在发音过程中有开始、中间和结束等状态，状态之间遵循一定的概率进行跳转。

每个状态会产生一个观察值（声学特征），这个观察值的概率分布可以用GMM （高斯混合模型）等来描述

语音识别系统：声学模型

声学模型将声学特征映射到发音单元，为后续的语言模型提供可能的语音单元候选



DNN-HMM 框架

DNN（深度神经网络）也应用到了语音识别的声学建模上，主要是替换了GMM-HMM模型中的GMM模型，上端仍然是HMM模型建模状态转移。

语音识别系统：语言模型

输入法的自动补全：南-> 京-> 大 -> 学

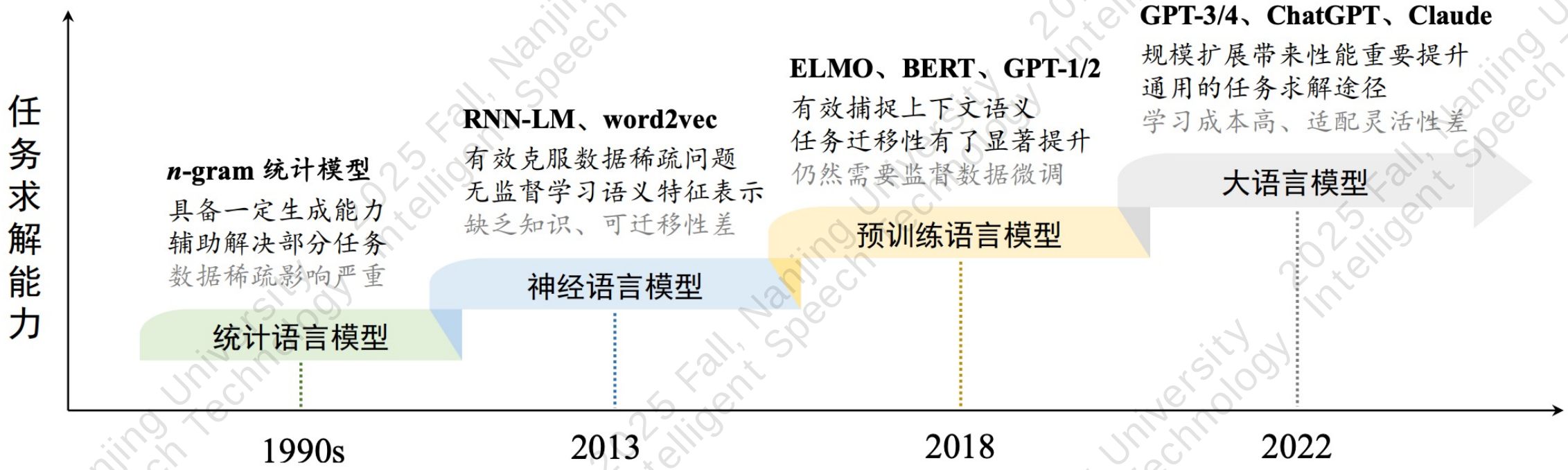
除了GPT 之外，大家日常都会接触到语言模型的应用，例如输入法，或者网页搜索的联想功能。

核心任务：基于大量文本数据的统计规律，根据当前历史，预测下一个词的概率

我很开心跟大家分享语音相关的只是。

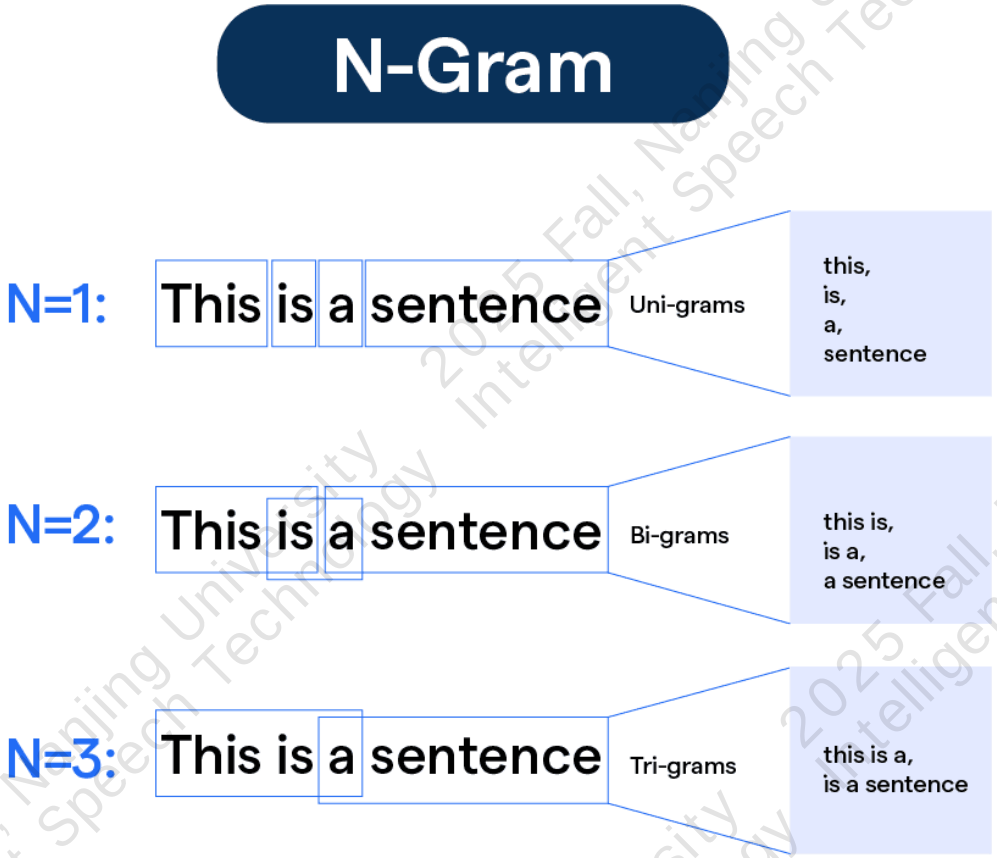
文档中的自动纠错

语音识别系统：语言模型



语言模型除了作为语音识别系统之外的一部分外，单独使用也展现了非凡的能力

语音识别系统：语言模型



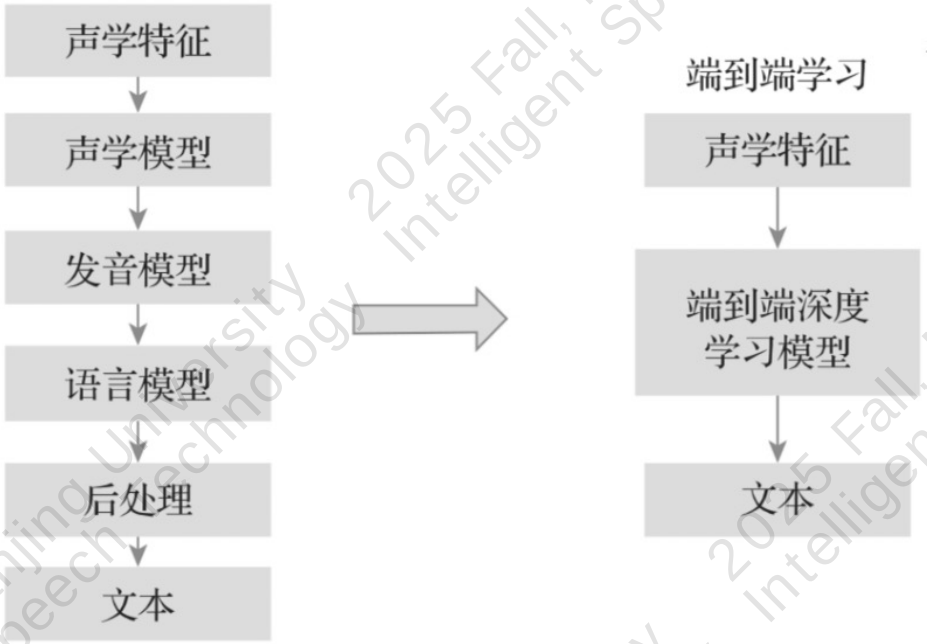
N-gram 是一种基于统计语言模型的方法，用于预测文本中的下一个单词或字符。

- Unigram 只考虑单个单词的概率，比如在一个文本语料库中，单词 “a” 出现的概率 P (“a”)。计算方式是单词 “a” 出现的次数除以语料库中单词的总数。
- Bigram 考虑两个连续单词的组合概率，例如 P (“a sentence”)，计算方法是 “a sentence” 这个单词对出现的次数除以 “a” 这个单词出现的次数。

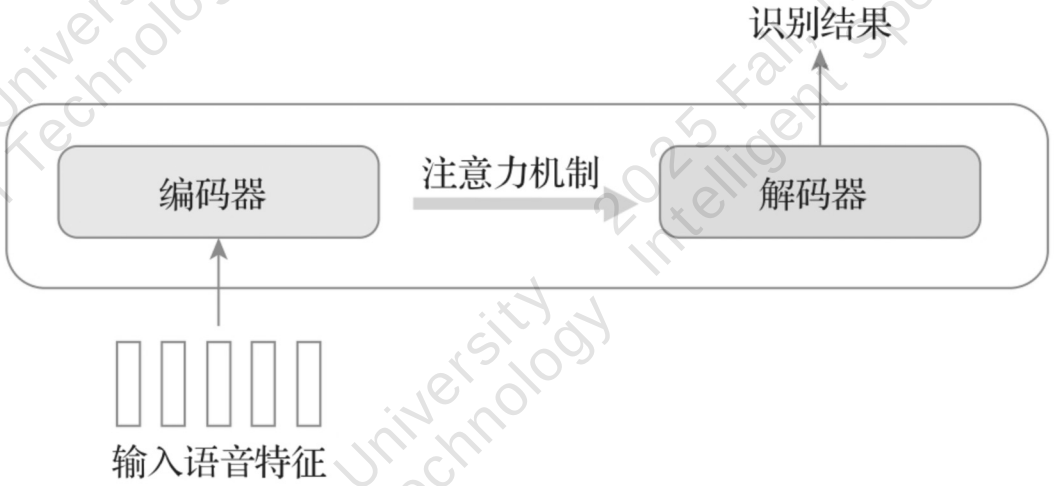
预测过程：在 bigram 模型下，如果当前单词是 “a”，我们可以查看所有以 “a” 开头的 bigram（如 “a cat”，“a dog”，“a sentence” 等），根据它们在语料库中的概率来选择下一个最有可能出现的单词。

端到端语音识别系统

传统语音识别系统



基于编码器-解码器的端到端语音识别系统



弱模块化，直接构造一个端到端的模型，输入声学特征，输出文本

技术发展史

参数合成

模拟声带激励和声道调制

拼接合成

单元选择方法，基于模板库选择发音片段拼接

统计模型

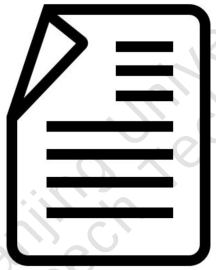
划分发音单元，为每一个发音单元构建传统统计模型如HMM

深度学习

采用神经网络替代传统的统计模型

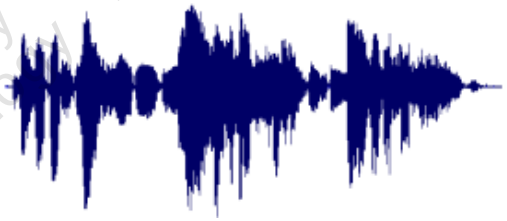
语音合成

输入文本 → 文本前端 → 算法后端 → 合成语音



- 文本正则化
- 分词
- 注音(多音字)
- …

- 拼接合成
- 统计模型
- 神经网络
- …



语音合成：文本前端

文本：欢迎南京大学 2025 级的同学们
正则化：欢迎南京大学二零二五级的同学们
分词：欢迎 / 南京大学 / 二零二五 / 级 / 的 / 同学们
注音：Huan2 ying2 Nan2 jing1 da4 xue2 er4 ling2 er4 wu3 ji2 de tong2 xue2 men5

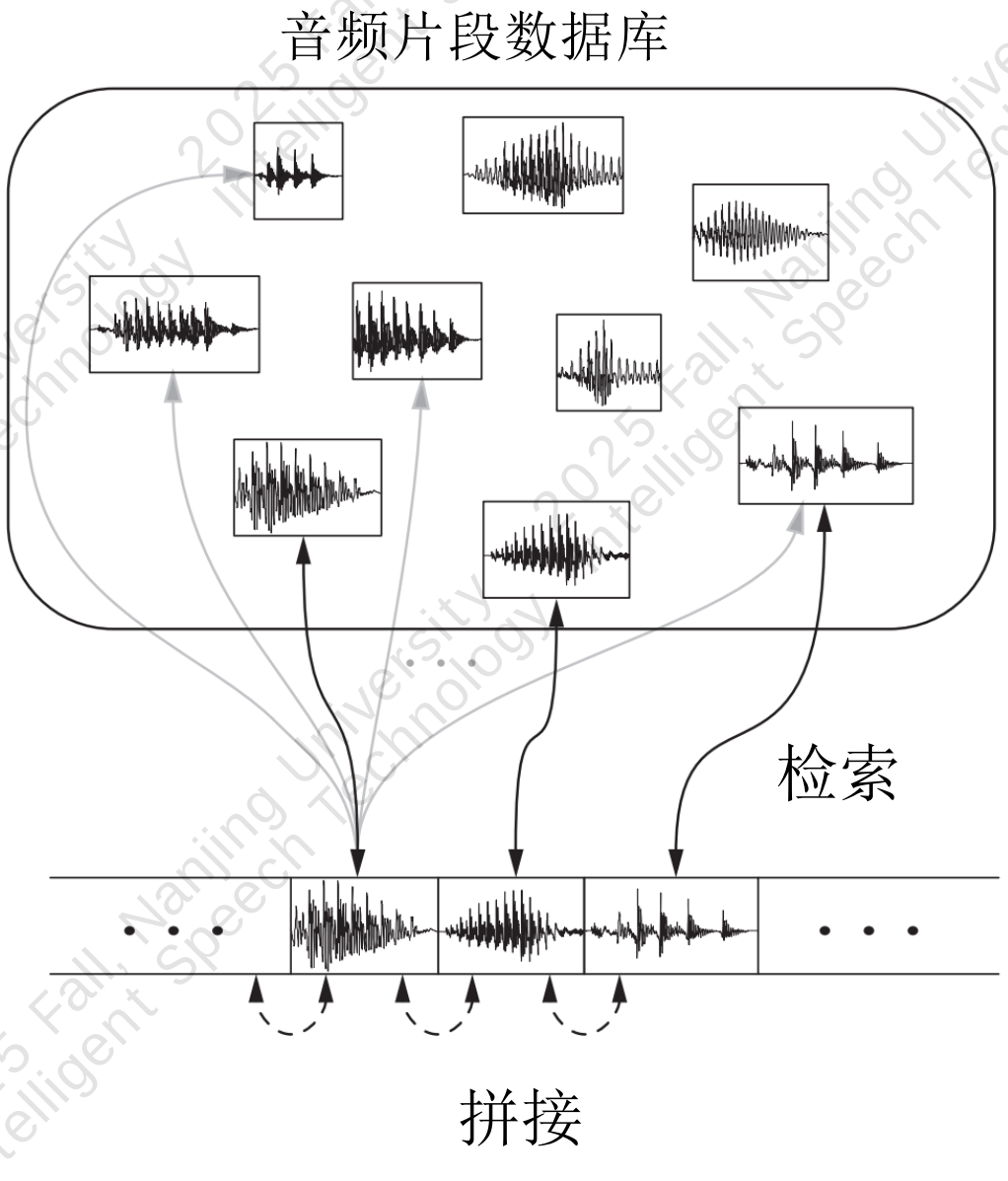
语音合成：算法后端

拼接法

1. 从录制的数据库中选择合适的语音片段
2. 拼接各个语音片段形成最后的合成语音

例如：希望合成“南京大学”，从数据库进行检索相应片段

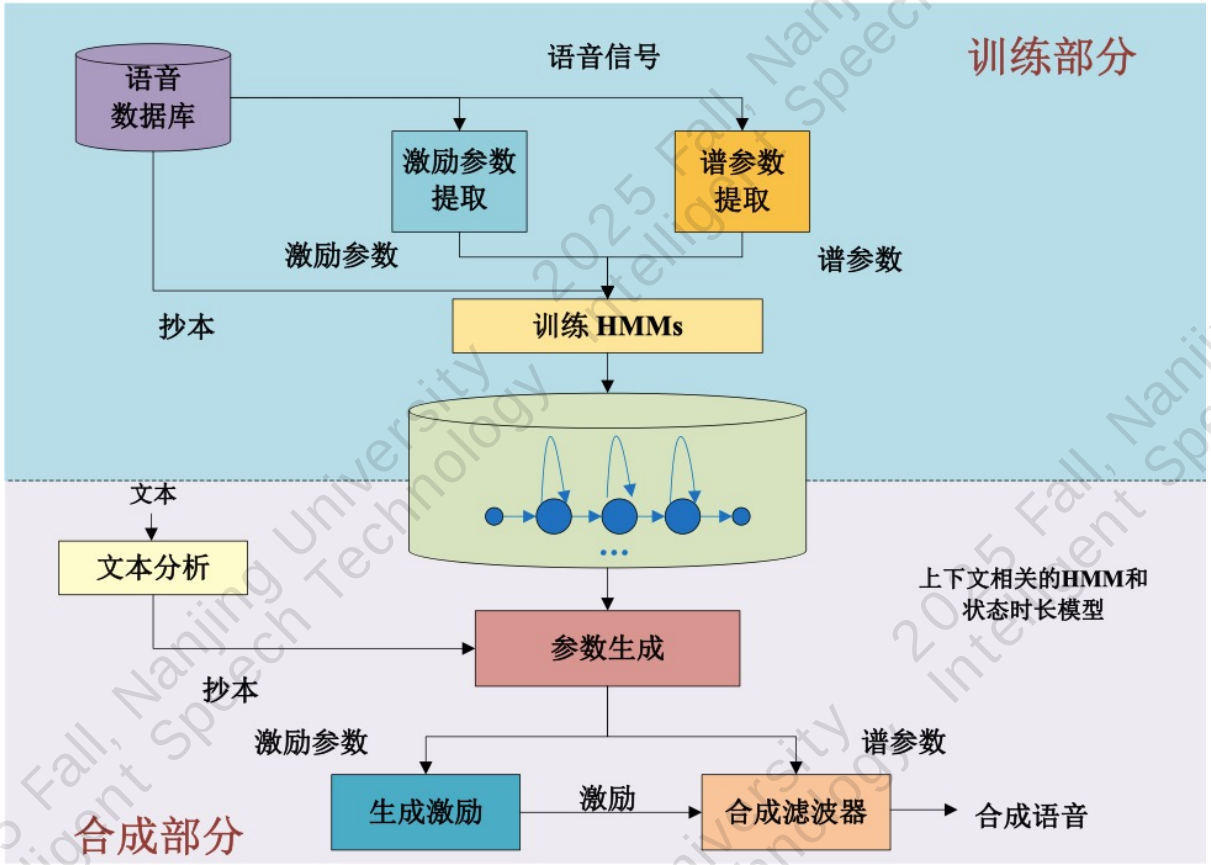
- 南方有个姑娘叫小芳
- 北京天安门
- 大学生活真美好



拼接合成示例：网络鬼畜视频

语音合成：算法后端

传统参数模型



基于HMM统计参数的语音合成

训练过程：建立文本到语音参数（如谱参数MFCC和激励参数基频信号等）的映射模型。

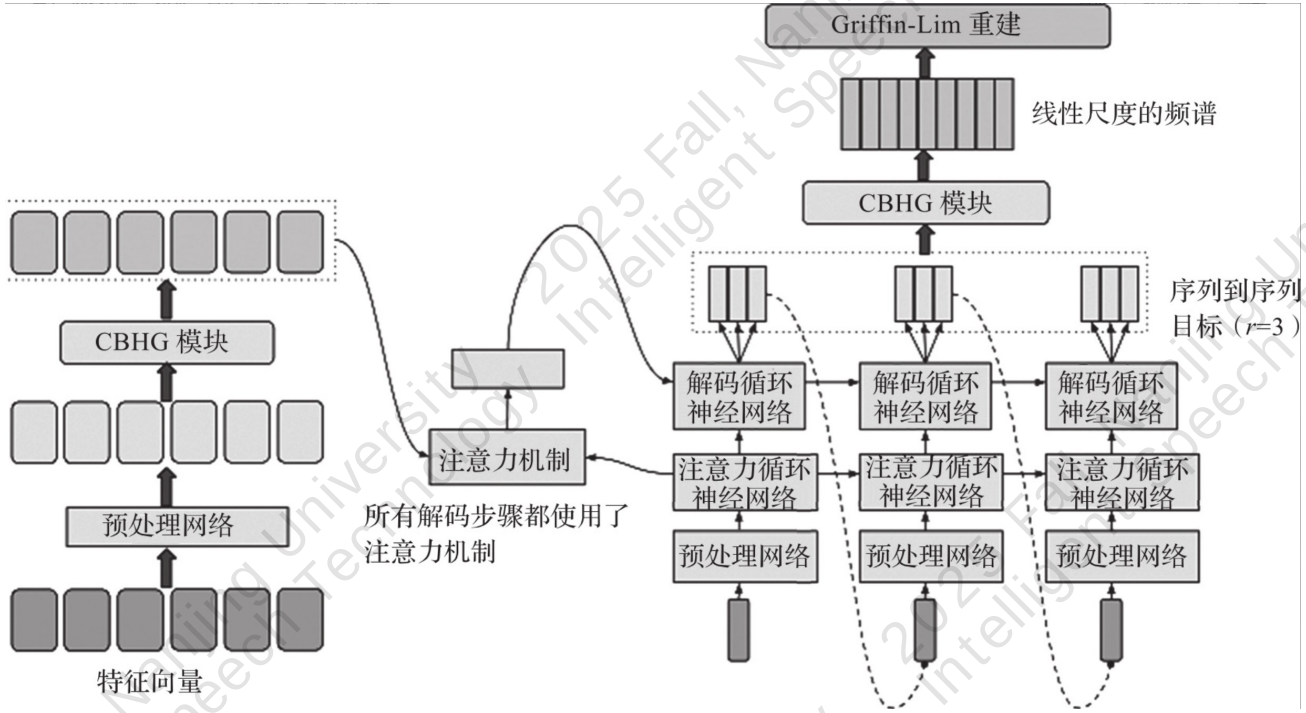
合成过程：对于接收到输入文本的HMM模型，通过参数生成算法，生成语音参数，通过逆梅尔频率倒谱变换等方法将其还原为频域信号，再经过傅里叶逆变换等操作得到时域信号。

参数合成示例

霍金采用的DECtalk 是典型的参数合成方法

语音合成：算法后端

深度端到端模型



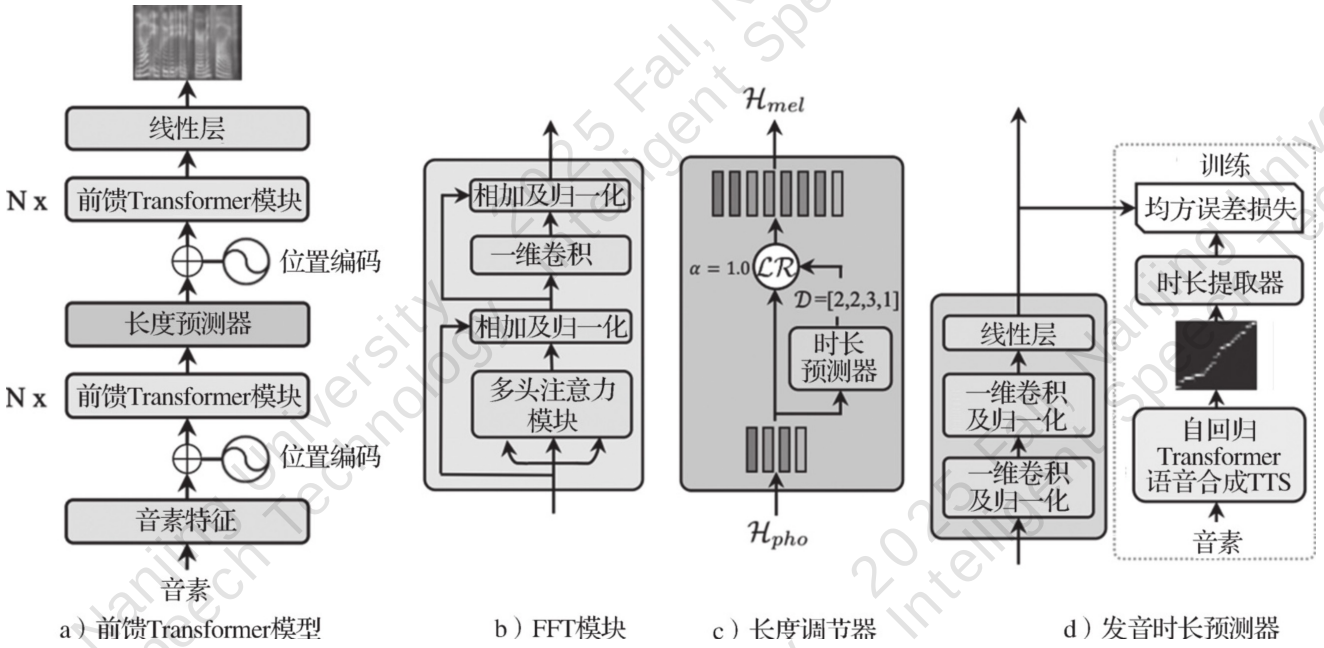
基于Tacotron 模型的语音合成，Google, 2016

- 编码器：负责将文本编码为中间特征向量
- 解码器：将中间特征解码为声学特征序列。
- Griffin-lim声码器：将预测出的梅尔谱转换成语音波形

由于发音时长全依赖自回归方法进行推测，合成音频中吞字漏字的现象很严重

语音合成：算法后端

深度端到端模型



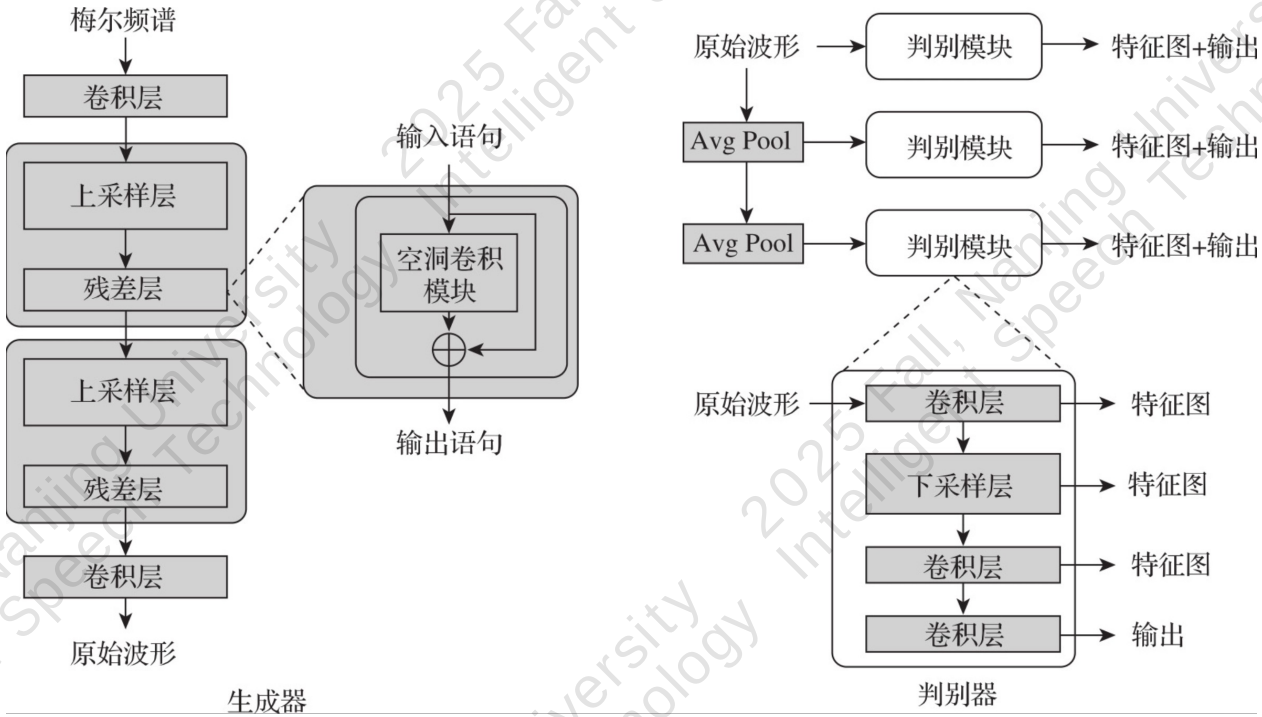
基于FastSpeech模型的语音合成，微软，2018

高效的端到端语音合成模型，能够快速生成高质量语音

- 发音时长预测器：预测每个音素的持续时长
- 长度调节器：调整音素序列与梅尔频谱序列之间的长度关系，通过预测每个音素的持续时间来对齐文本和语音特征

语音合成：算法后端

高质量的声码器：Mel到音频波形重构



基于GAN（生成对抗网络）的声码器

声码器

将梅尔频谱转换为实际可听的语音波形，把声学特征还原为原始的语音信号，使语音能够通过扬声器等设备输出。

基于生成对抗网络（GAN）的声码器由生成器（Generator）和判别器（Discriminator）两部分组成，通过两者之间的对抗训练来生成高质量的语音。

评测标准：主观评测，听音打分

分级指标	含义描述
5 分	整体语音自然流畅、发音清晰、易于听懂，无法区分是模型合成的语音还是真人发音
4 分	听起来比较清晰自然、发音清晰，虽然没有严重的韵律错误，但是能明显辨别出是否属于模型合成
3 分	音质可接受，语音不太流畅，有部分发音错误或者不正常的韵律起伏，有部分发音不清晰
2 分	音质比较差，语音不流畅，有简单堆积的音节，部分词语发音不清晰，听起来理解困难
1 分	听起来是明显的机器音，不流畅，难以理解，只能听懂只言片语

客观评测标准：与目标语音的音色相似度等

零样本语音合成 （音色克隆）

豆包声音复刻大模型

- 基于豆包语音大模型打造的超轻量级音色定制方案
- 开放环境中录制秒级别录音即可极速拥有专属定制音色
- 广泛应用于视频配音、数字人驱动、语音助手、在线教育等场景

5秒
超低录制成本

秒级别
快速复刻

高还原
真人音色特点

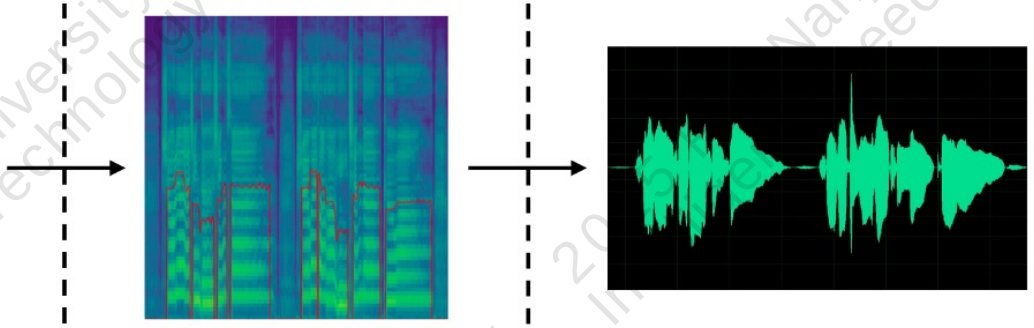
歌声合成

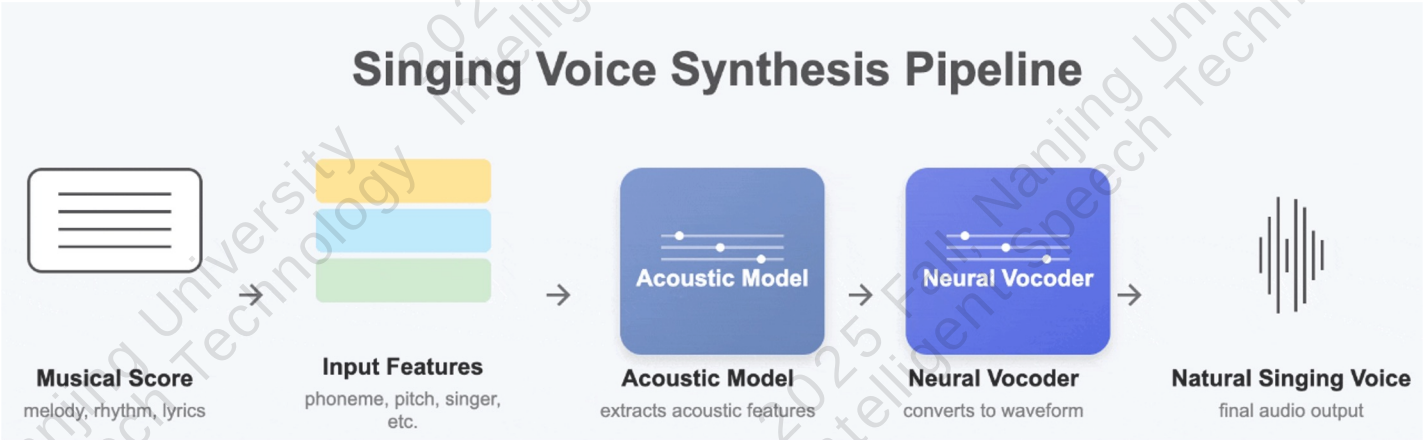
曲谱

歌词

雨淋湿了天空 灰得更讲究
你说你不懂 为何在这时牵手
我晒干了沉默 悔得很冲动
就算这是做错 也只是怕错过
在一起叫梦 分开了叫痛
是不是说 没有做完的梦最痛
迷路的后果 我能承受
……

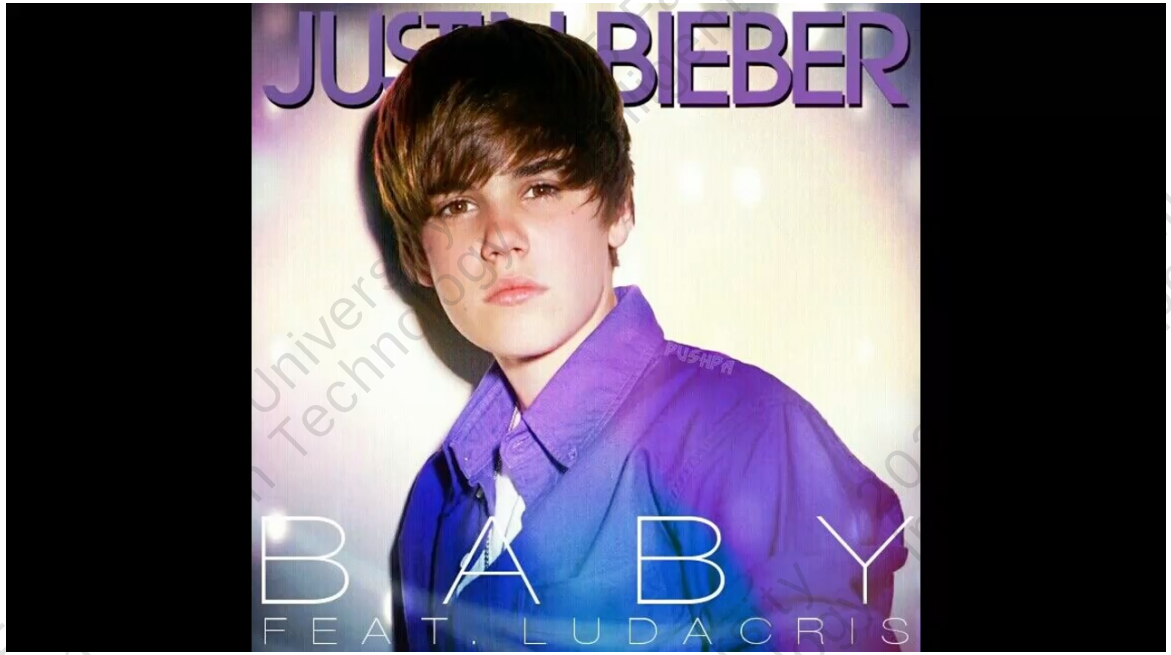
频谱





歌曲生成

同时生成人声+伴奏



歌曲生成

同时生成人声+伴奏

《解决了当前高学历人才无法进行rap创作的问题》

2024-12-17

♡ 1190 ◯ 89

♥ 3734 ☆ 1177 ◯ 1548

学历这么高的大模型是学不会说唱的

2024-12-16

♡ 2548 ◯ 201

准备投哪个期刊

2024-12-16

♡ 376 ◯ 172

作者列表看着都有股制作人名单的味了

2024-12-20

♡ 9 ◯ 1

中国有嘻哈

标题也太有意思了

01-09

♡ 2 ◯ 1

感觉他们做的应该也很开心哈哈

01-10

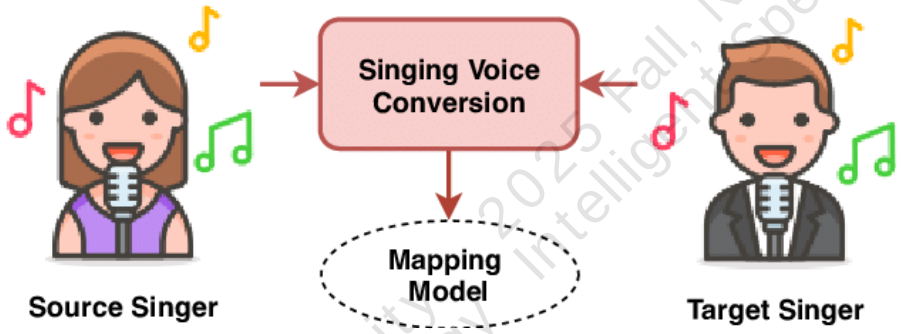
语音转换

语音转换：将源语音转换为目标说话人的声音

语音转换

语音转换：将源语音转换为目标说话人的声音（抖音音色）

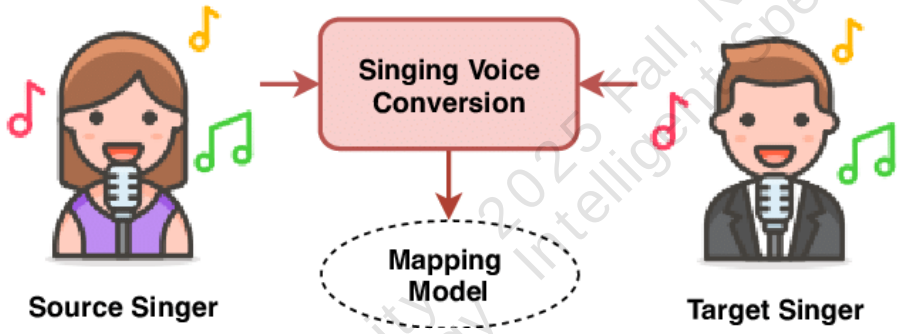
歌声转换



相对于普通的语音转换

更加注重对**韵律，情感**的建模

歌声转换



相对于普通的语音转换

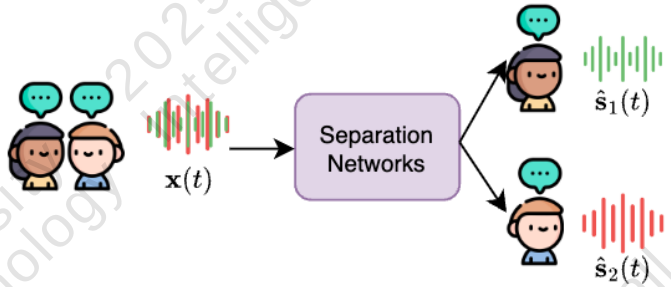
更加注重对**韵律，情感**的建模

鸡尾酒会问题

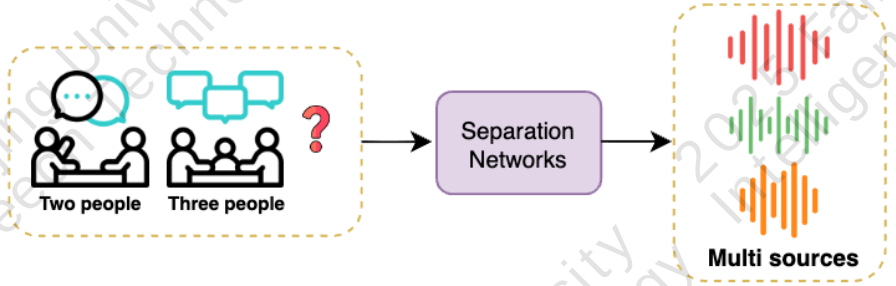
在嘈杂环境中，如何专注于某一个声音源？

人类天然有选择性注意力听觉

将重叠的语音分离成单人的声音



(a) Known number of speakers



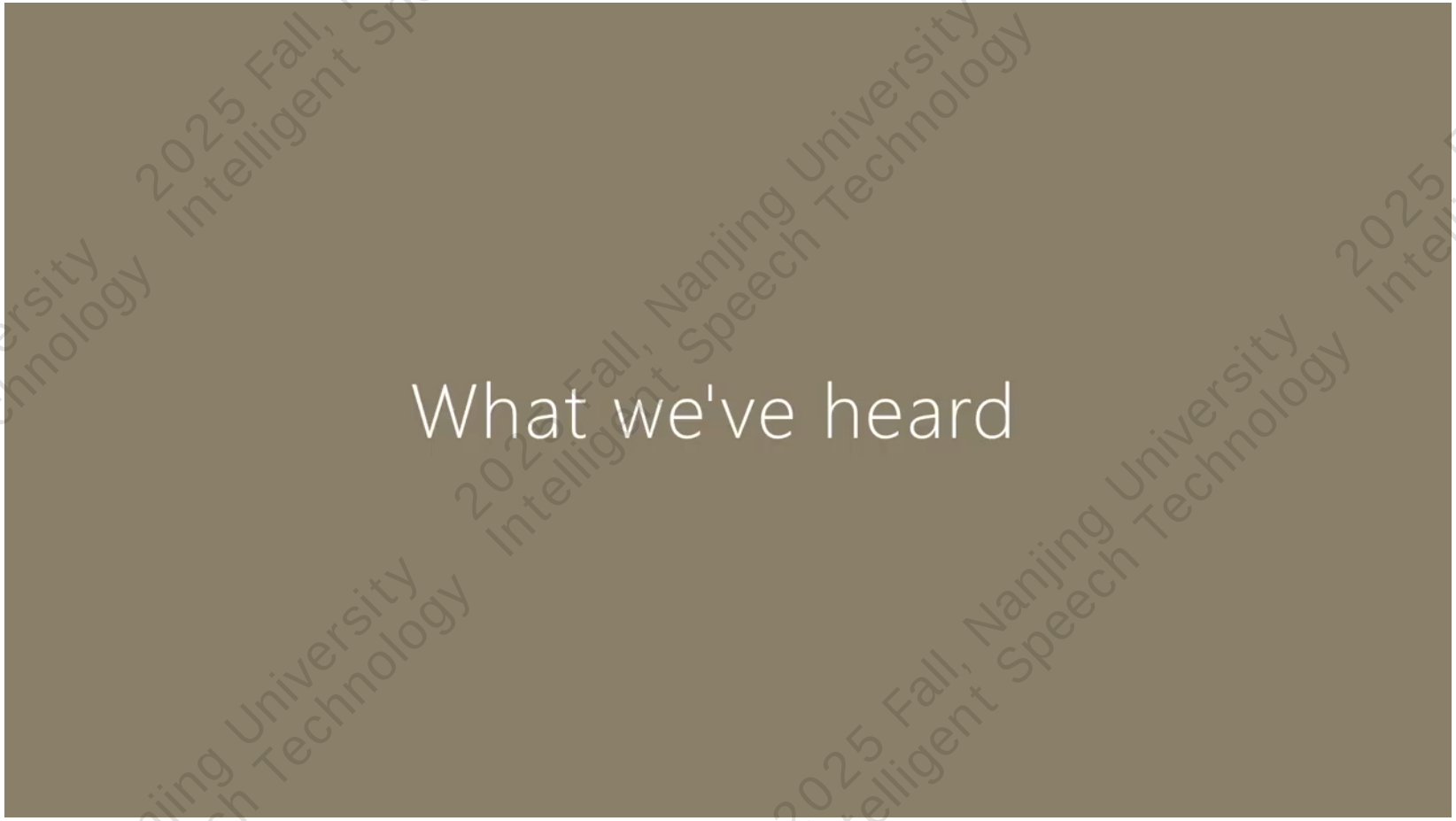
(a) Unknown number of speakers

Core Problem:

$$\mathbf{x} = \sum_{i=1}^{C} \mathbf{s}_i + \mathbf{n}$$

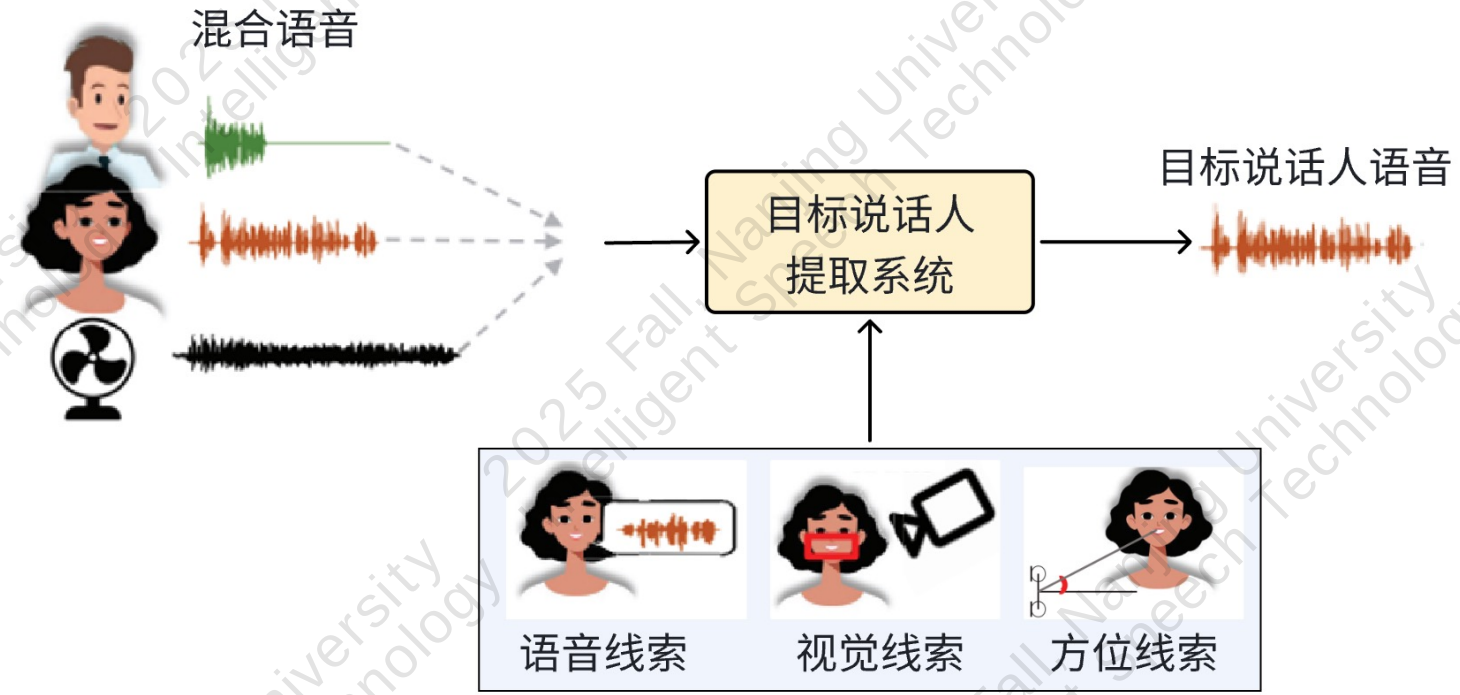
语音分离

将重叠的语音分离成单人的声音，与目标人声音提取不同之处在于没有额外的目标参考信号，因此也经常被叫做“盲源分离”。



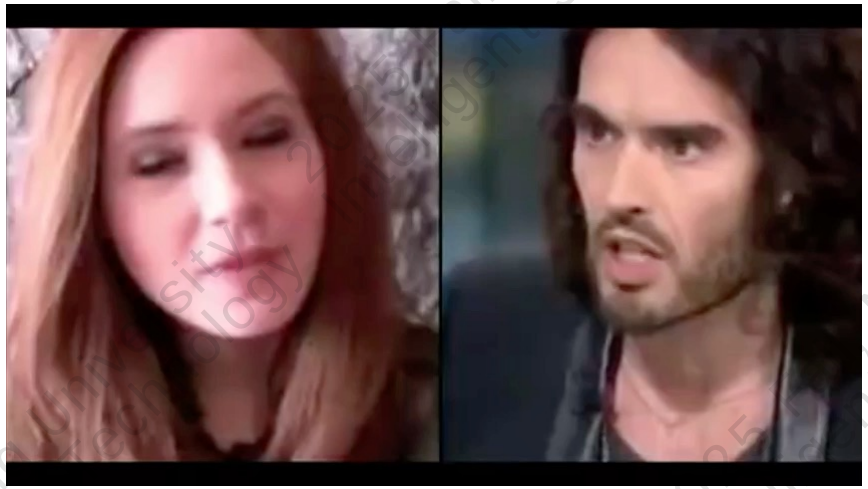
目标说话人提取

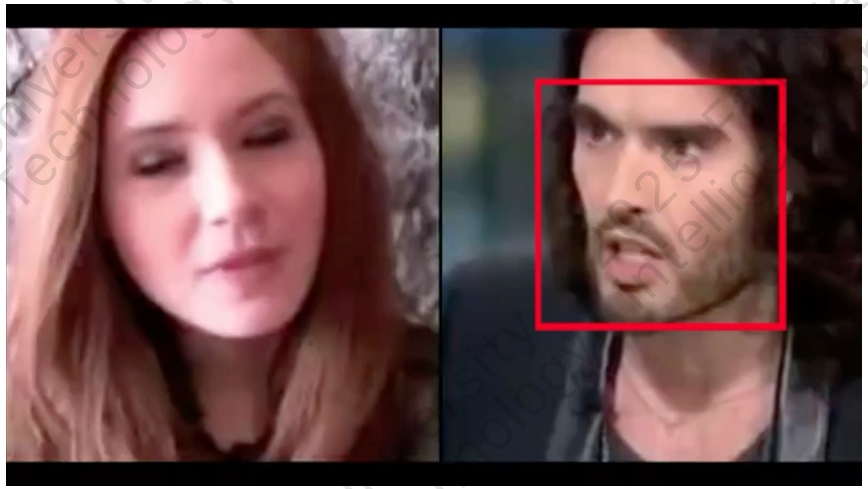
如果我们只关注某个特定人呢？



目标说话人提取

如果我们只关注某个特定人呢？





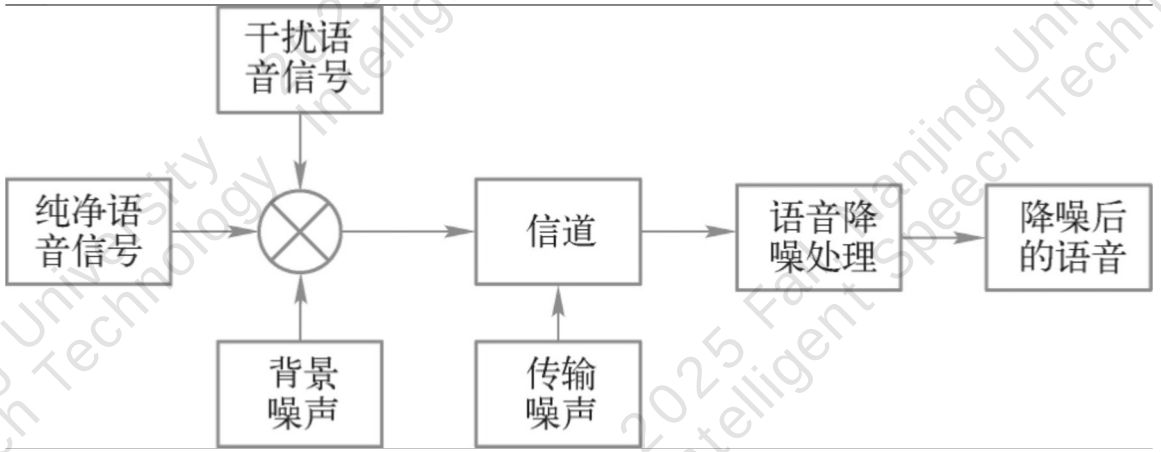
- 目标更清晰：有“谁”的先验（参考音频/声纹），直取目标；分离是盲的，后续还要配对。
- 稳健不漂移：多人重叠/噪声/混响下更少身份互换与切换。
- 下游更顺滑：输出即目标身份，免“成分指派”，ASR/检索/验证更稳。

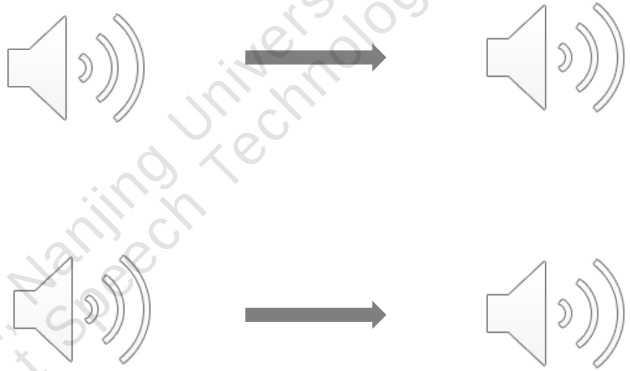
语音增强

常见的噪声种类

加性噪声
- 周期性噪声：例如钟表的滴答声
- 脉冲噪声：例如突然的一声锣响
- 宽带噪声：例如街道上混合的嘈杂声
- 语音干扰噪声：其他的人声

非加性噪声
- 混响电路噪声：例如房间内的回声
- 传输网络噪声：例如网络信号不好时的语音失真

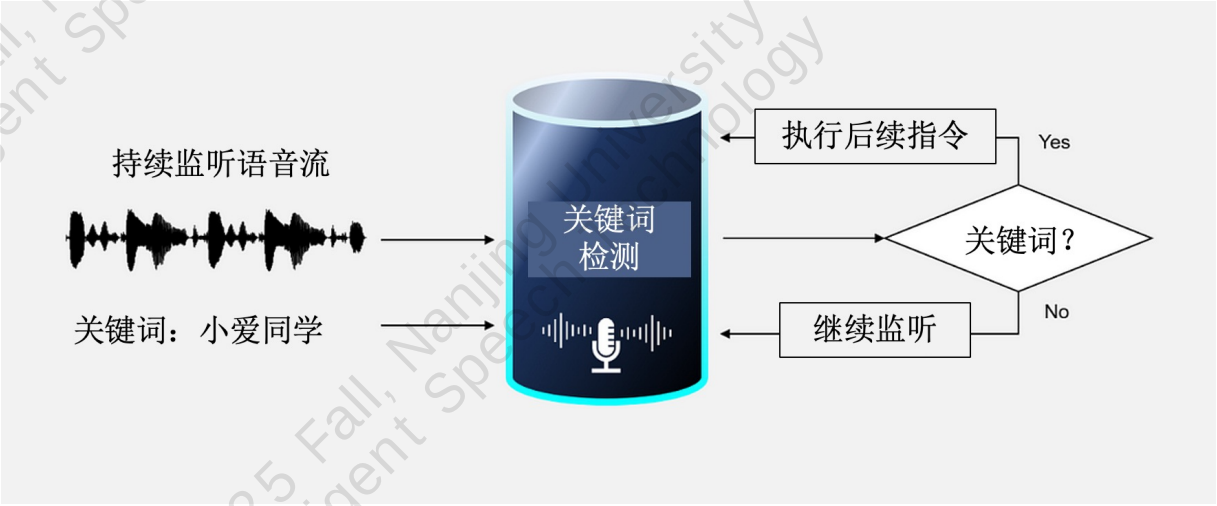




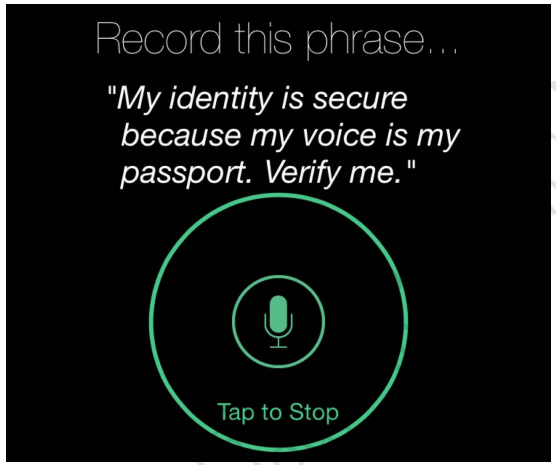
语音唤醒

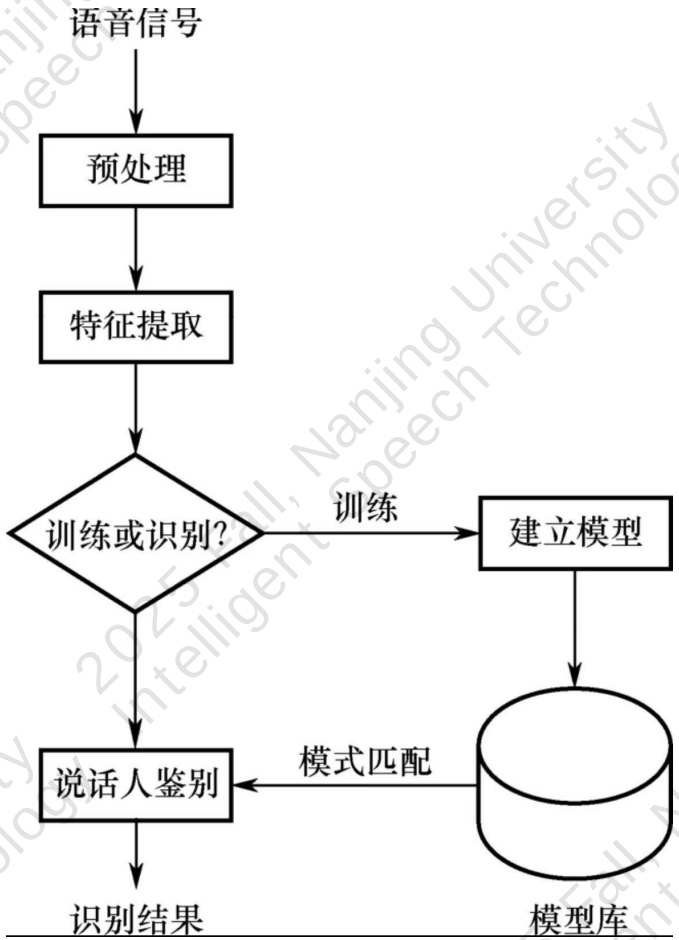
工作流程：

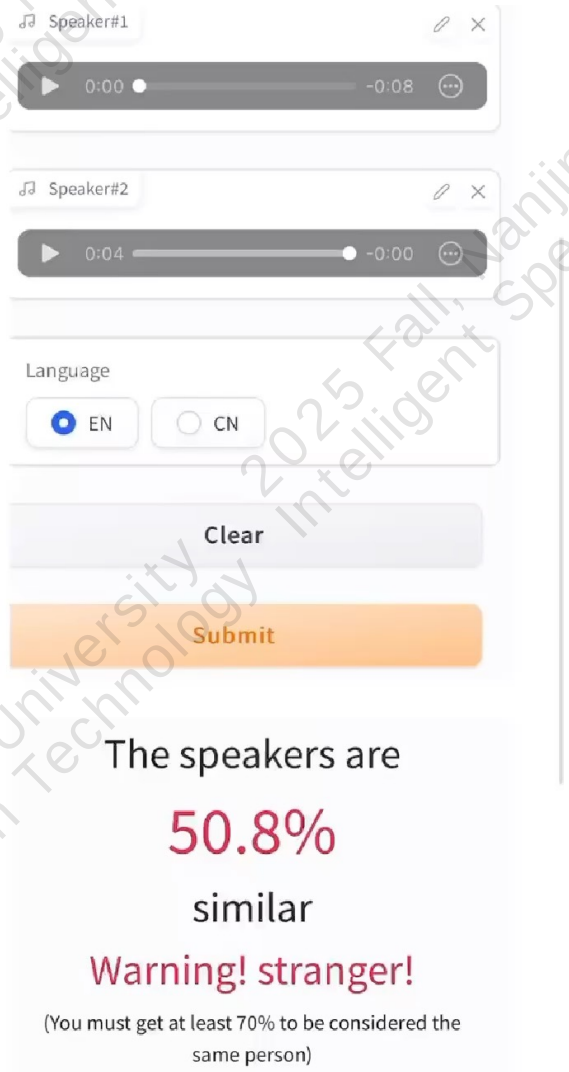
- 持续监听环境声音
- 检测是否含预设唤醒词，若有则唤醒设备



声纹识别



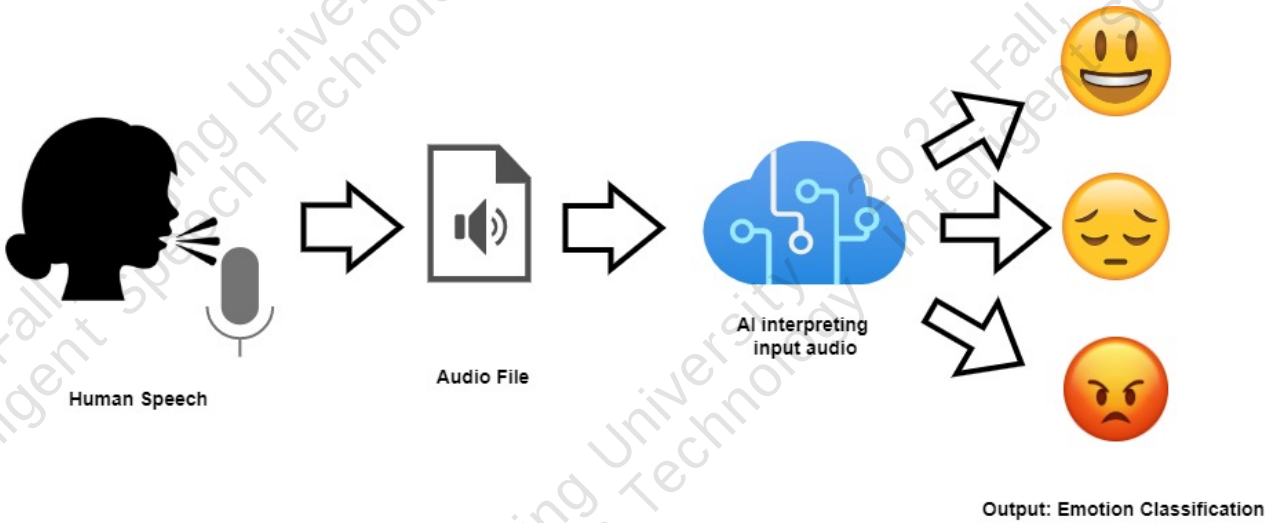




听懂你的喜怒哀乐

近年来，情感研究愈发重要

- 人们对于计算机的需求已不是仅仅满足功能需求，更需要照顾到人的情 感需求
- 计算心理学的兴起
- 人机交互系统的跨越式发展，希望更有温度的人机交互



陶建华, 陈俊杰, 李永伟. 语音情感识别综述[J]. Journal of Signal Processing, 2023, 39(4).

情感表示：从离散到连续

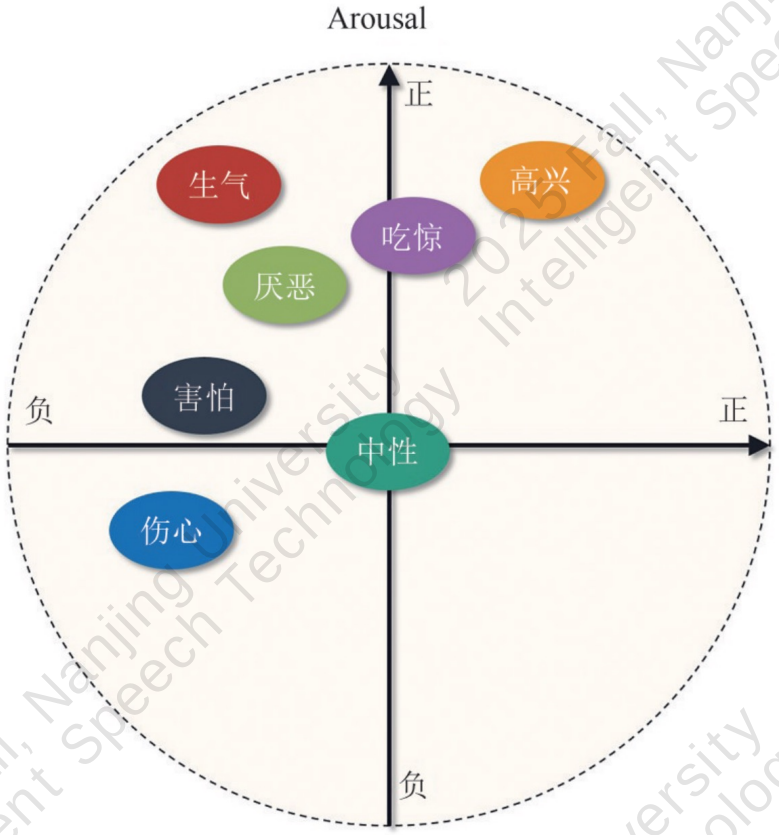


图 1　Valence-Arousal 维度情感空间

离散表示：例如 喜、怒、哀、中性四分类

连续表示：效价度-激活度（Valence-Arousal）模型

- **激活度A** 表示个体的神经激活水平
- **效价度V** 表示个体情感状态的积极或消极性

相对于离散表示：

不仅有定性的描述，还多了定量的分析，能够反映更加细微的情感变化

陶建华, 陈俊杰, 李永伟. 语音情感识别综述[J]. Journal of Signal Processing, 2023, 39(4).

语音翻译

语音到语音，适合没有书面文字的语言

世界上有超过6000种语言，但是只有不到3000种有自己的文字

目录

CONTENTS

智能科学与技术学院
School of Intelligence Science and Technology

统一范式：序列到序列

序列到序列建模：将源序列转换成目标序列

任务类型	输入内容	输出内容
对话系统	用户的问题或话题	合适的回复内容
机器翻译	一种语言的文字	另一种语言的文字
语音识别	一段语音信号	语音中人的说话内容
语音合成	一段文字	与文字对应的语音信号
语音问答	用户的语音提问	语音回复内容
语音转换	一种语音	转换后的语音 （如不同音色、口音等）

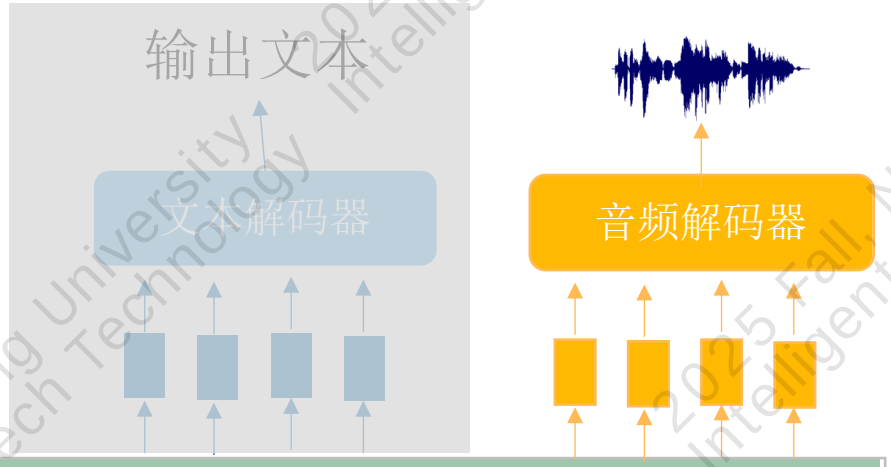
统一范式：序列到序列

文本到文本

- 机器翻译
- 问答系统
- 文本生成

语音到文本

- 语音识别
- 语音到文本翻译
- 语音内容描述



语音到语音

- 语音编辑
- 语音转换
- 语音到语音翻译

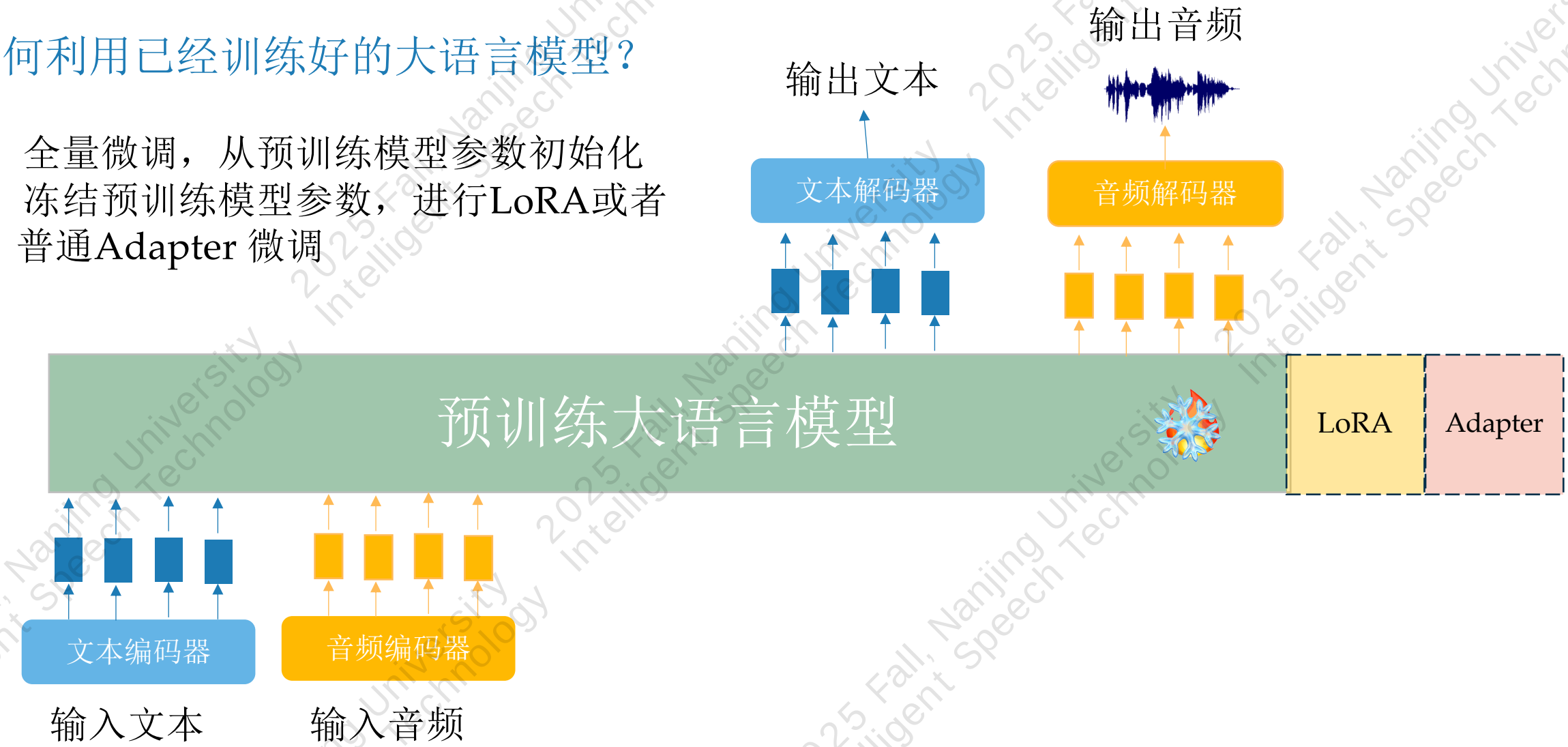
文本到语音

- 语音合成

统一范式：序列到序列

如何利用已经训练好的大语言模型？

1. 全量微调，从预训练模型参数初始化
2. 冻结预训练模型参数，进行LoRA或者普通Adapter 微调



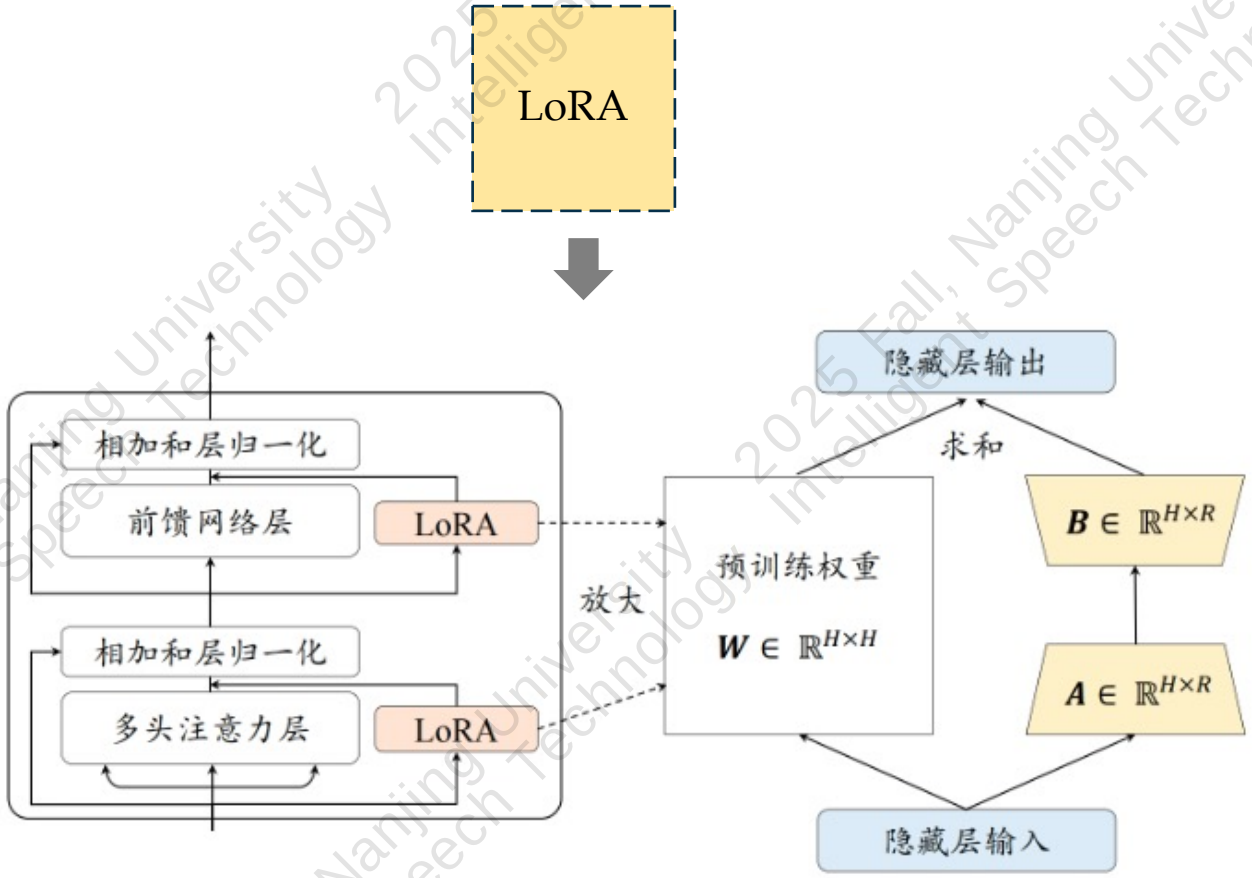
统一范式：序列到序列

高效微调方法

大语言模型的参数矩阵的维度通常很高，针对特定任务进行微调时，不需要调整全量参数。

LoRA（Low-Rank Adaptation, 低秩适配）

提出在预训练模型的参数矩阵上添加低秩分解矩阵来近似每层的参数更新，从而减少需要训练的参数量



以transformer 层为例的LoRA 示意图

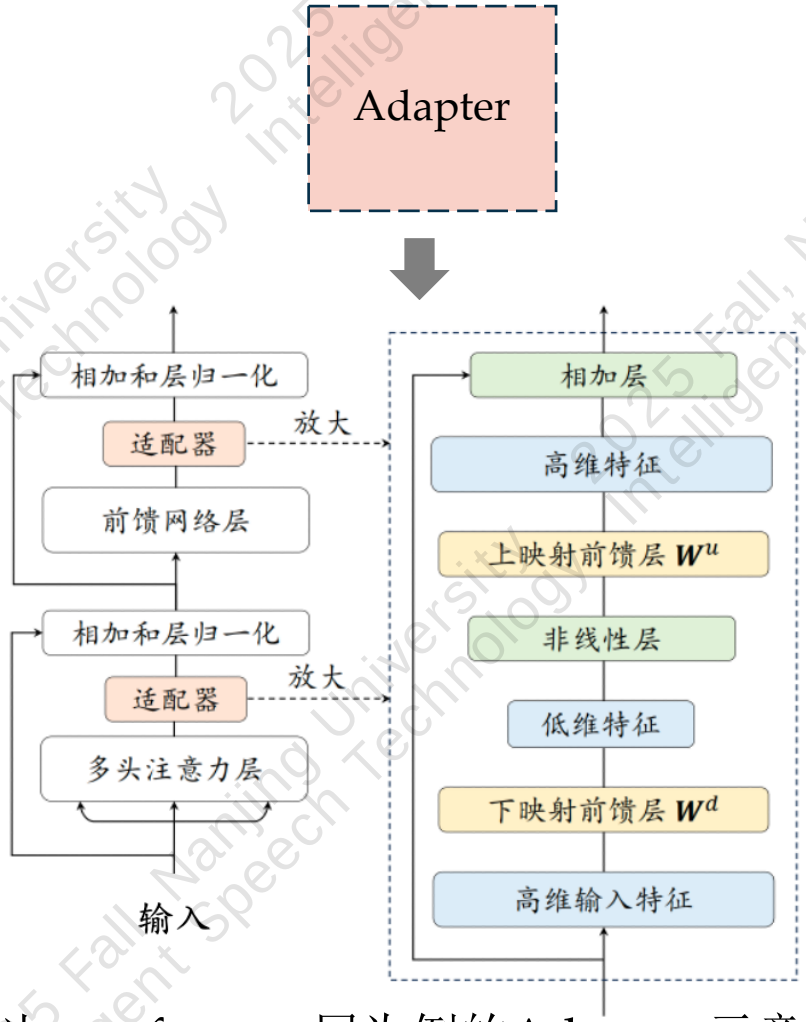
统一范式：序列到序列

高效微调方法

大语言模型的参数矩阵的维度通常很高，针对特定任务进行微调时，不需要调整全量参数。

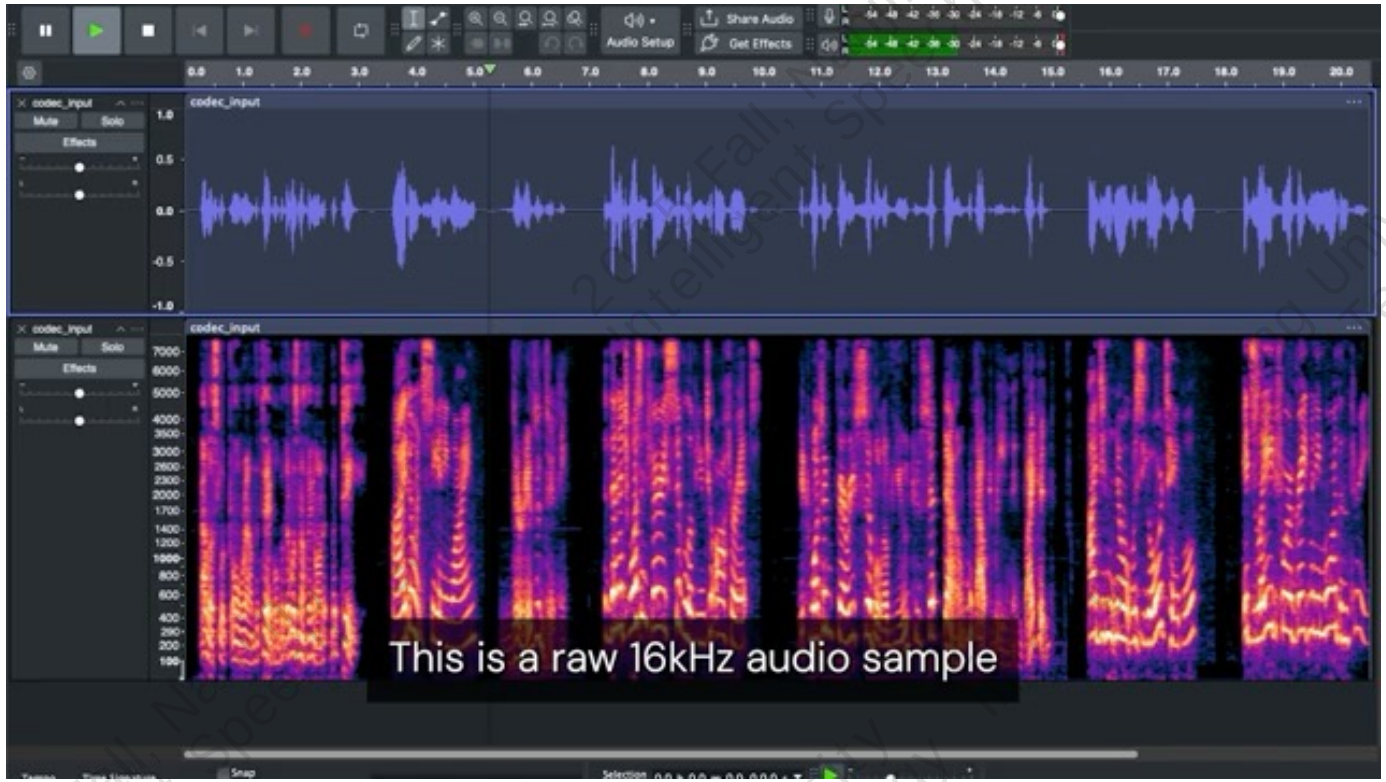
适配器微调（Adapter Tuning）

在模型层间引入小型神经网络模块（适配器），先将原始的特征向量压缩到较低维度，使用激活函数进行非线性变换，最后再将其恢复到原始维度。



以transformer 层为例的Adapter 示意图

核心问题：语音的高效编码



Codec 做了什么？

- **压缩冗余**：语音里相邻采样点、相邻帧高度相关，Codec 用预测、量化、熵编等减少重复信息。
- **保真优先**：人耳不敏感的部分（如某些高频/掩蔽区）更可压，人耳敏感的共振峰、F0 等被优先保留。
- **实时传输**：把连续波形切成小帧，编码成小包；更易抗丢包、重传、纠错。
- **可控延迟**：调"帧长/码率/缓冲"三要素，在音质、带宽与时延之间平衡。
- **鲁棒工具箱**：基于声学模型的降噪、回声/增益控制常与 Codec 协同工作。

核心问题：语音的高效编码

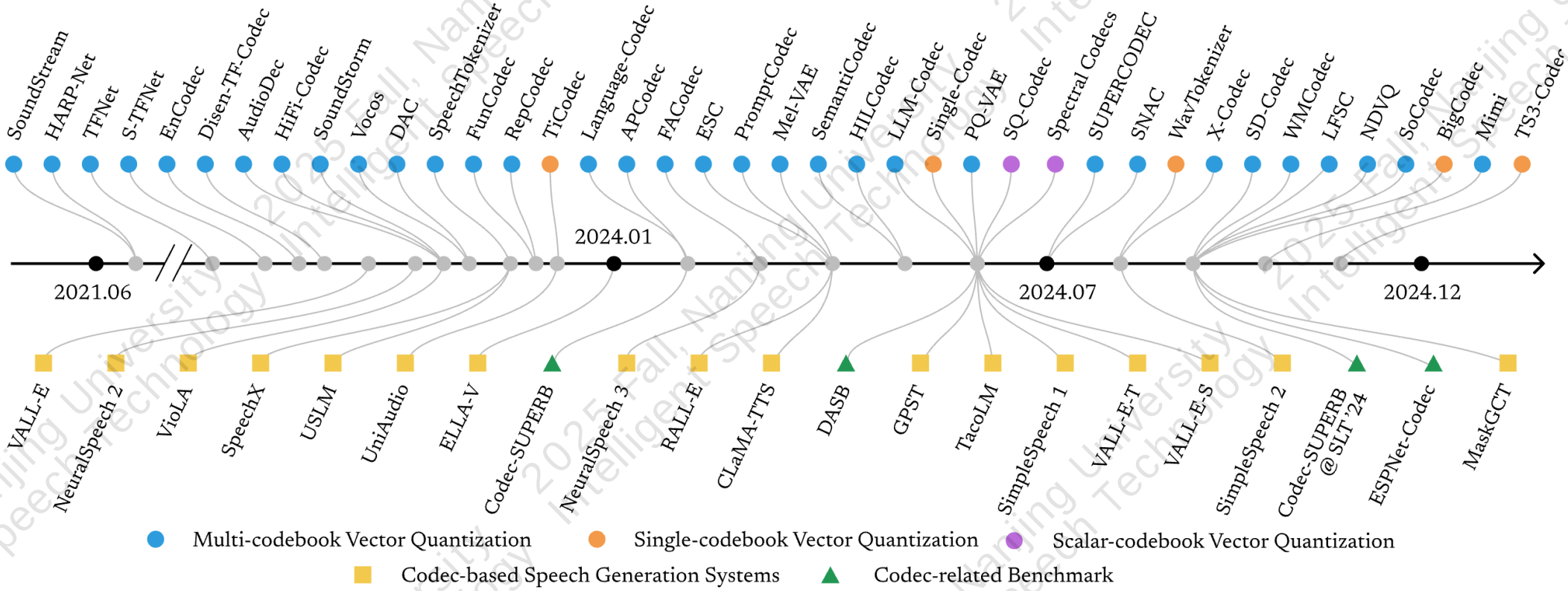


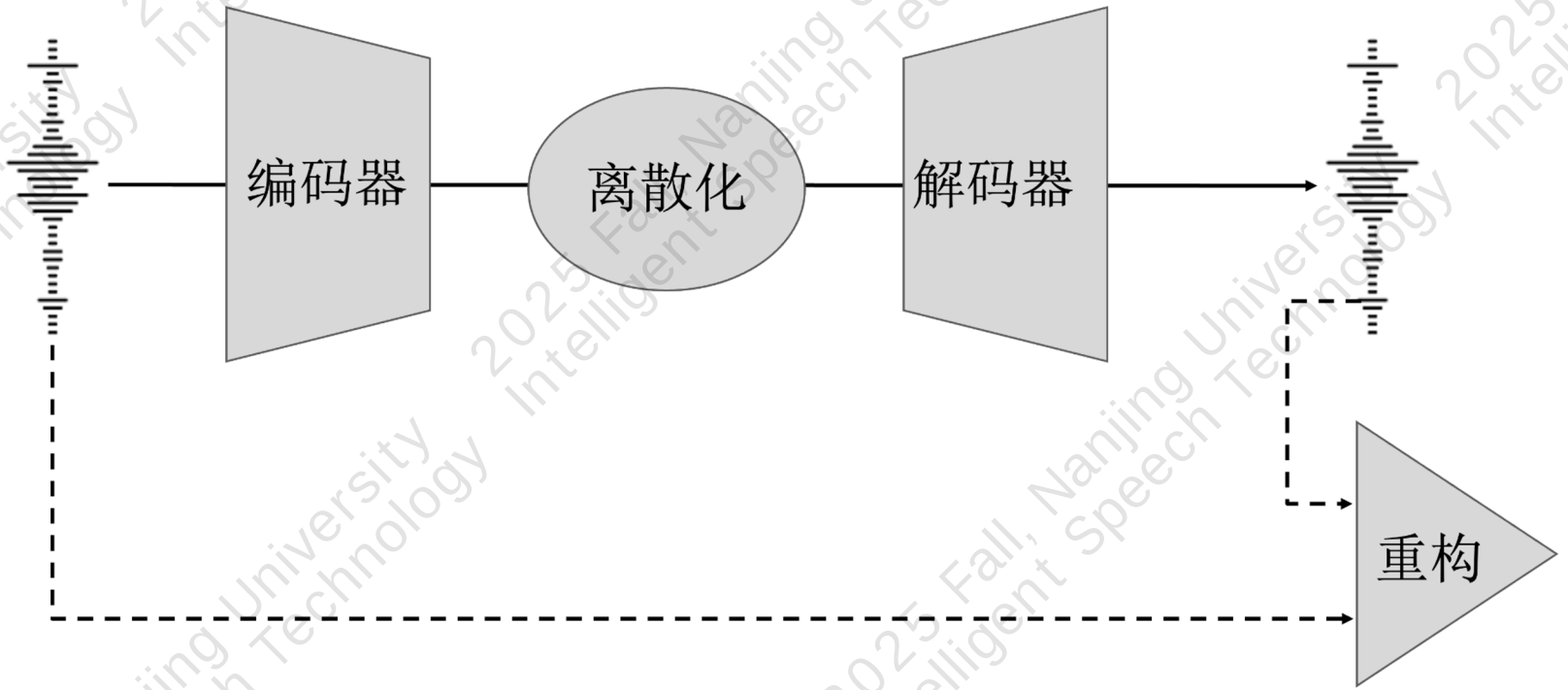
Fig. 2. Timeline of current different types of neural audio codec models and codec-based speech generation models.

图片来源：Chen X, Du J, Wu H, et al. CodecFake+: A Large-Scale Neural Audio Codec-Based Deepfake Speech Dataset[J]. arXiv preprint arXiv:2501.08238, 2025.

核心问题：语音的离散编码表示

基于信号重构的语音编码

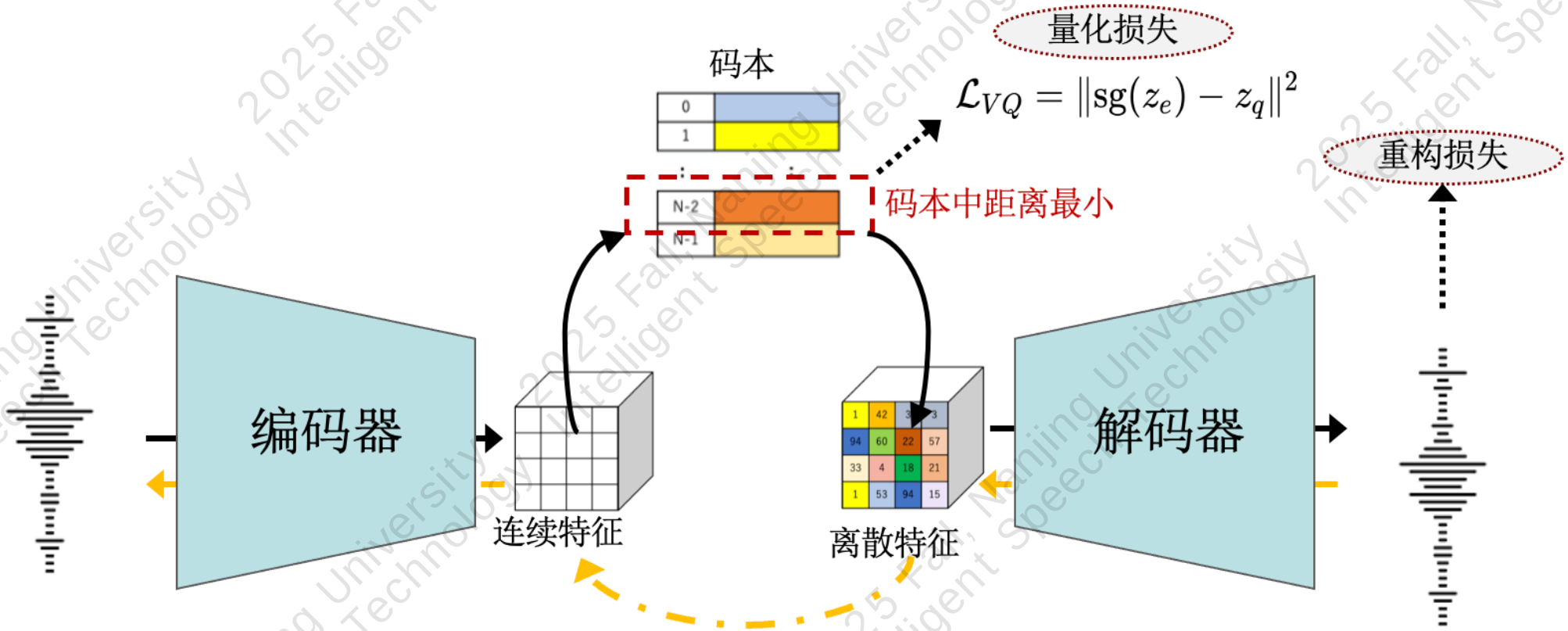
编码器输出特征为连续空间，如何像文本信号一样进行离散化？



核心问题：语音的离散编码表示

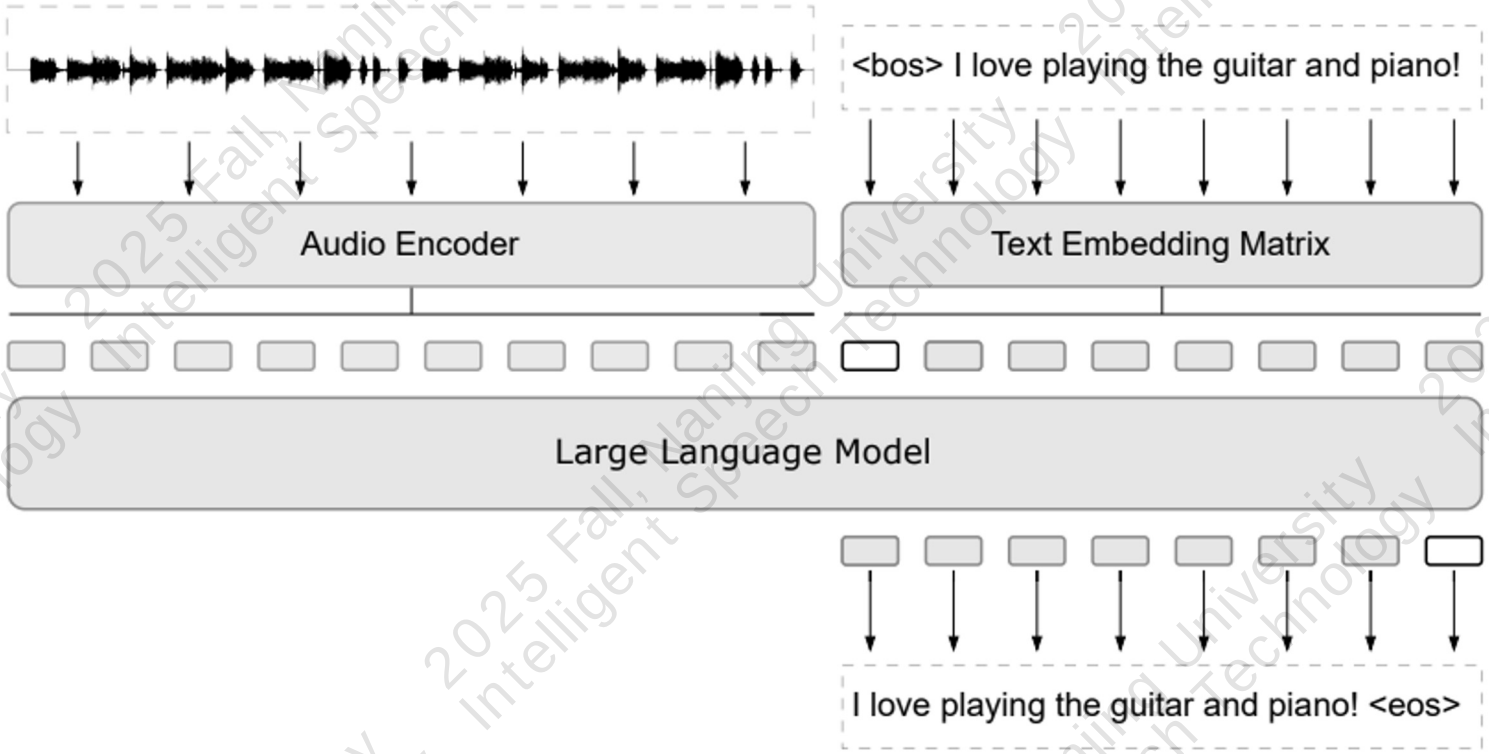
在线离散化方法：

矢量量化算法，设置一组码本，将当前的连续向量替换为距离最小的码本



例子：LLM重塑的语音识别新范式

输入语音

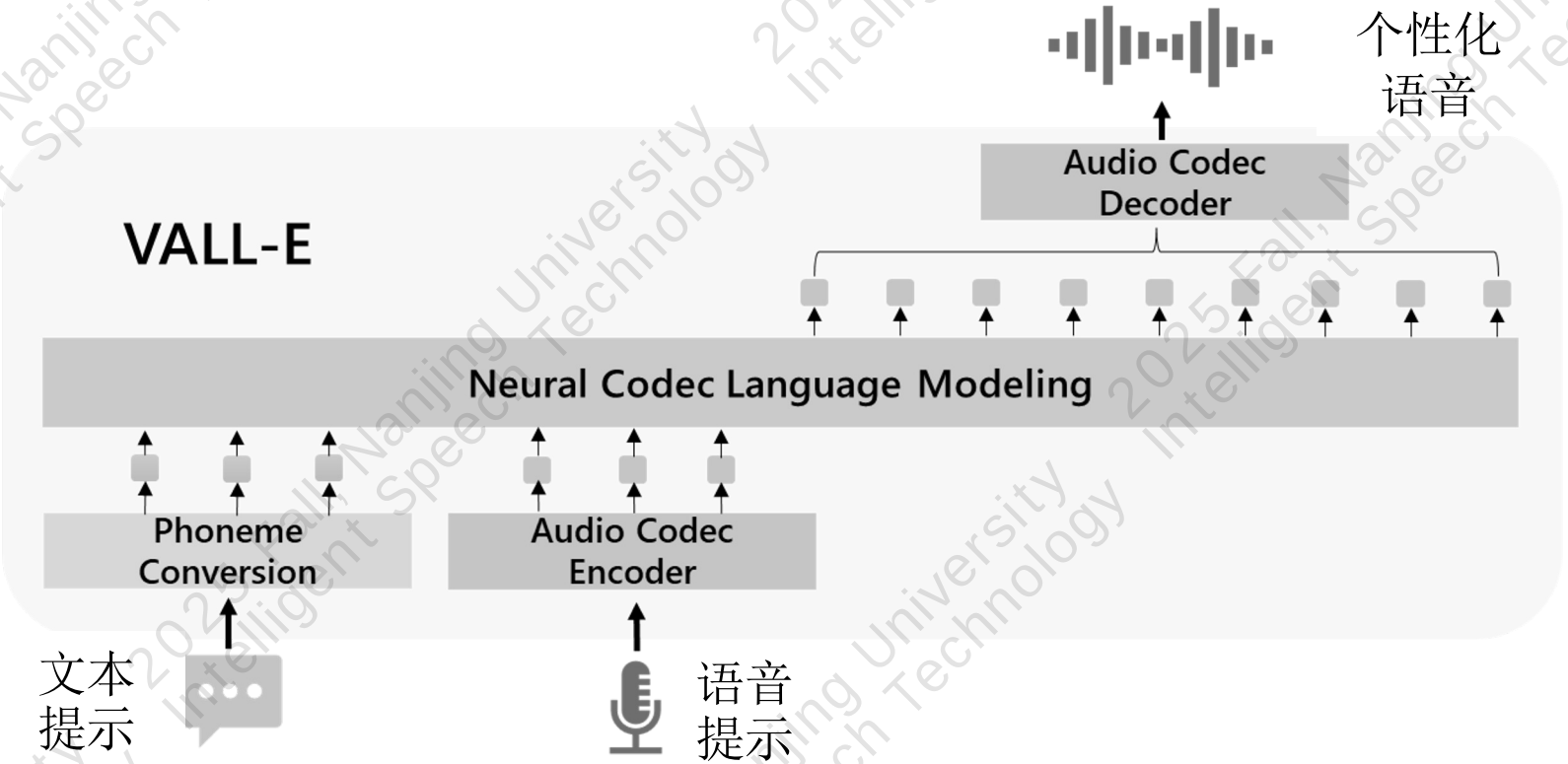


识别文本

拼接语音信号和目标文本，自回归方式预测文本的下一个词

例子：LLM重塑的语音生成新范式

- 音色克隆效果提升明显
- 跨语种合成容易实现



示例：音色克隆

中文原音

英文配音

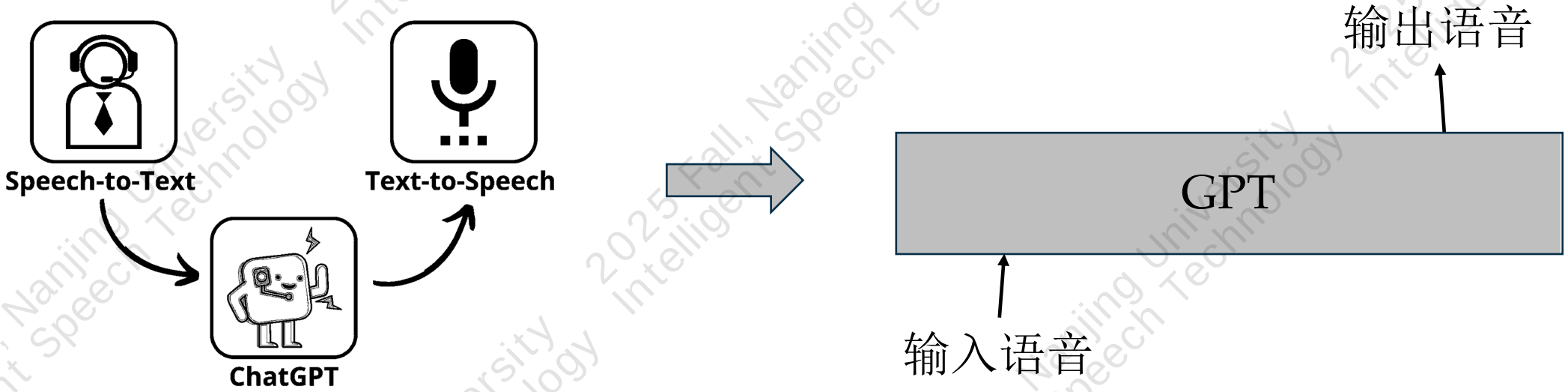
短时频配音，保持讲解人的

- 音色一致
- 语调一致
- 语速一致

来源：字节跳动SeedTTS

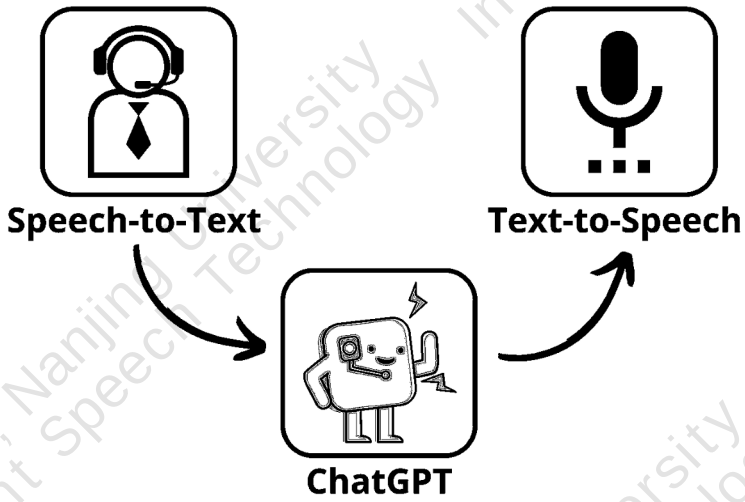
LLM重塑的端到端语音对话系统

LLM 的建模框架，实现文本、语音模态的融合，进而实现端到端的语音对话系统



LLM重塑的端到端语音对话系统

我们为什么需要端到端的对话系统？



各个模块的技术相对成熟，很容易打造一个“能用的”对话系统

如果想要一个“好用的”对话系统，劣势明显

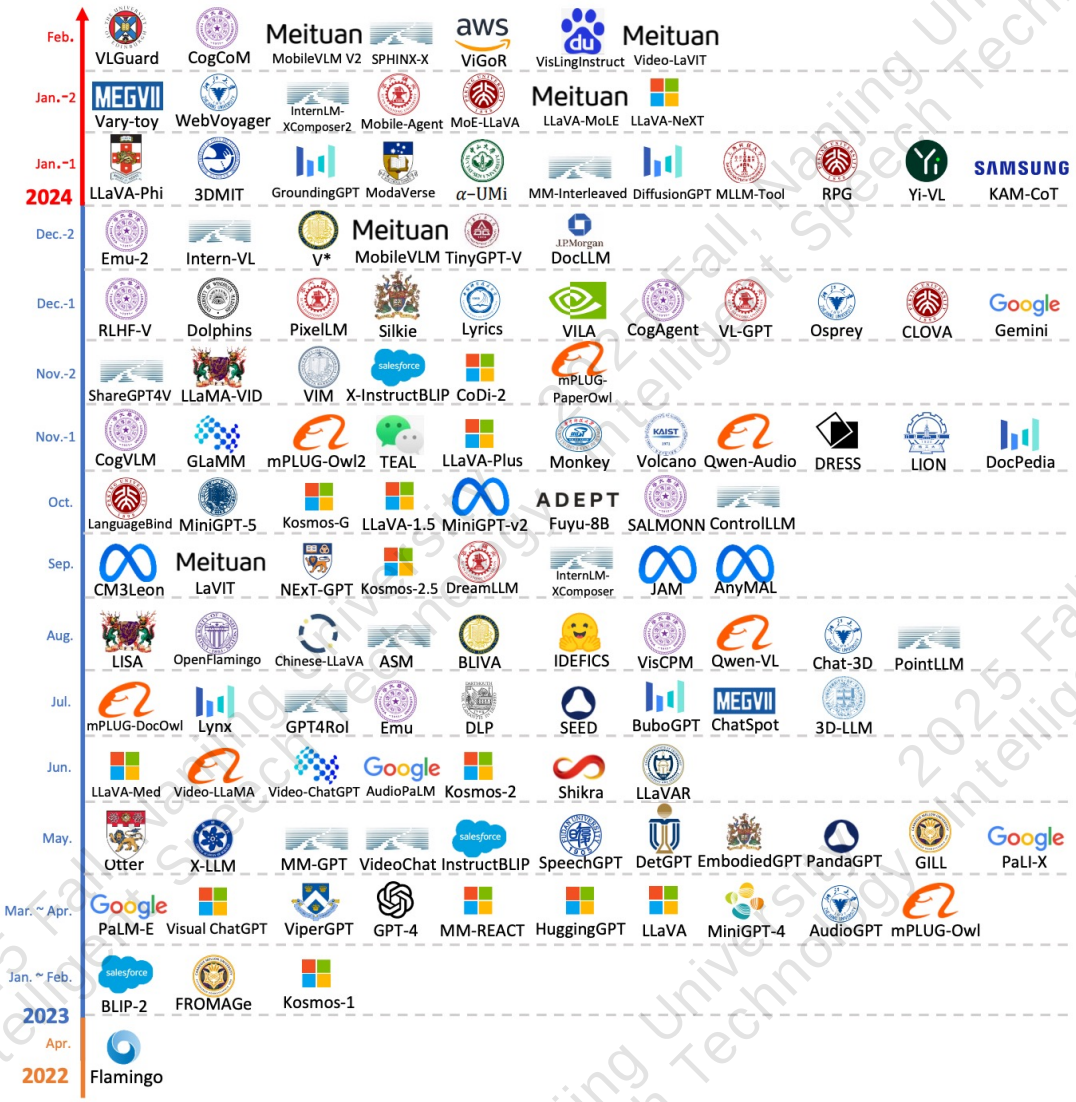
- 丢失信息：语速、情绪、音调、重音
- 不够真实：如无法进行有效打断
- 系统累积错误
- 系统时延较长

示例：实时对话

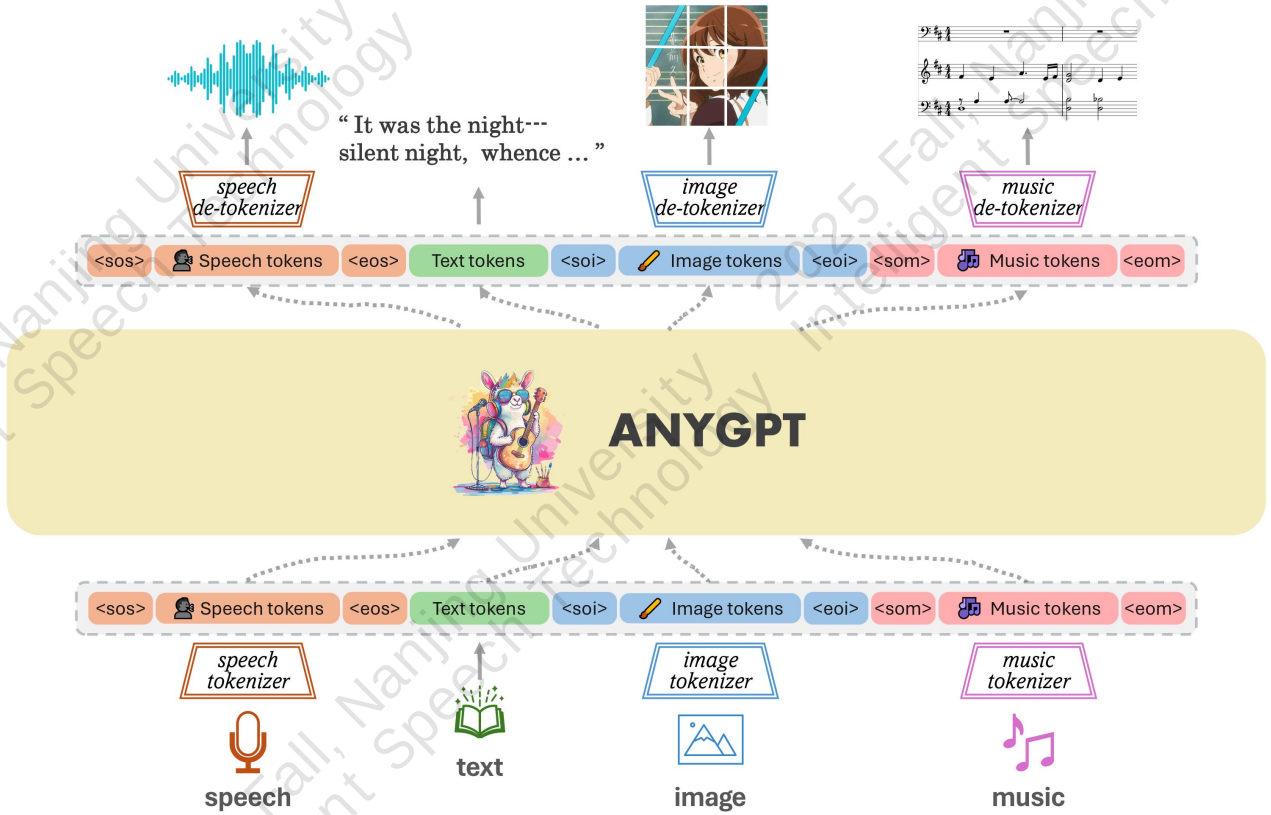
ChatGPT 支持口语教育，帮助学生锻炼英语表达能力和理解能力。并支持纠正语法错误、提供更地道的表达方式等。

- 地道的口语表达
- 实时的语音交互

多模态大模型



融合语音、图像、文本三种模态



AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling

示例：多模态对话系统

挑战：是否能完全依赖智能语音？

语音系统的鲁棒性问题

目录

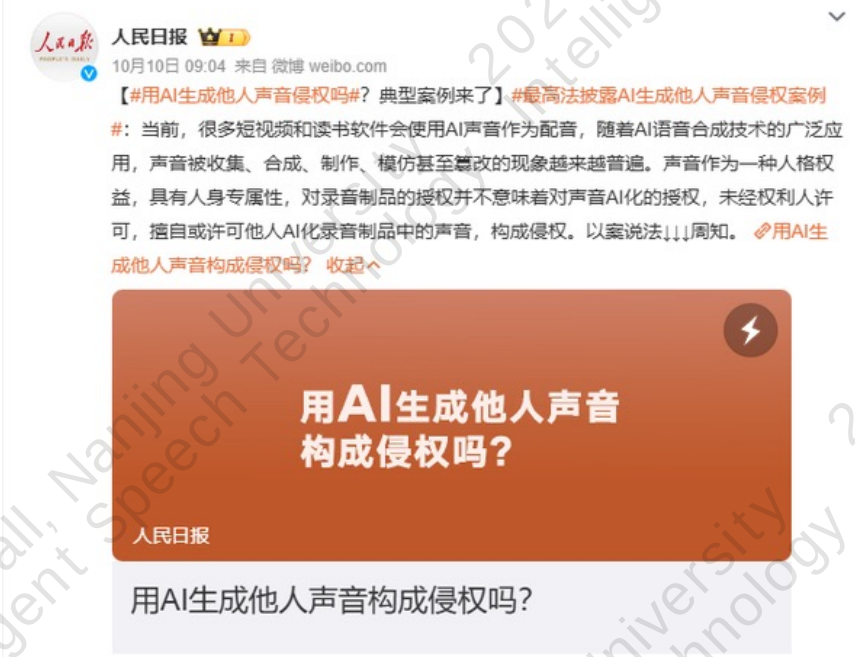
CONTENTS

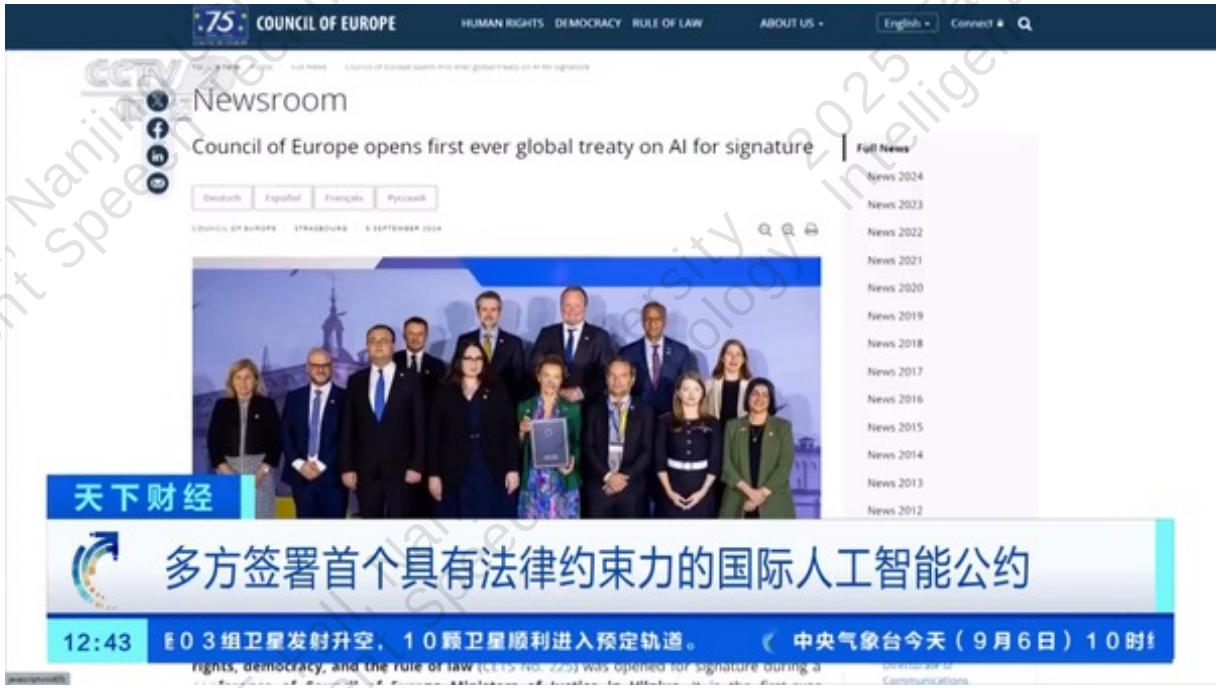
智能科学与技术学院
School of Intelligence Science and Technology

挑战：生成式AI的造假技术

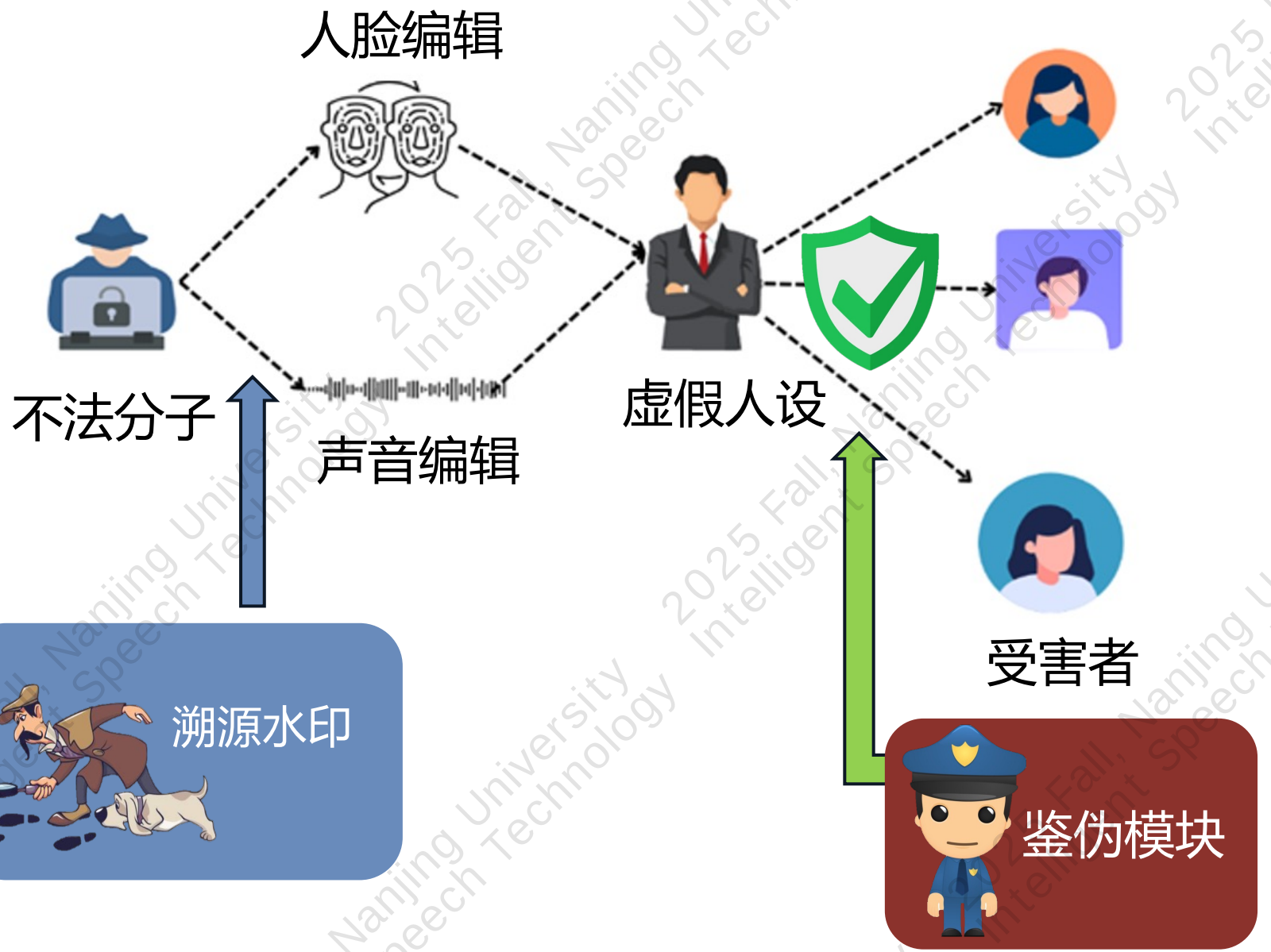
警惕 AI 语音技术的滥用

近期的“雷军骂网友”事件，“三只羊”录音伪造事件。。。





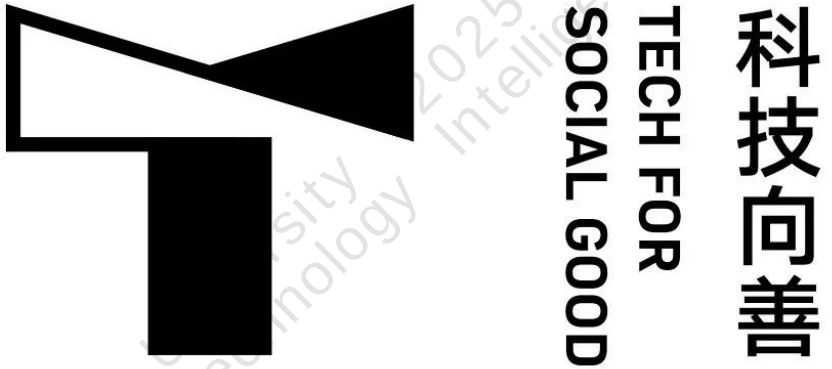
技术上：积极防御

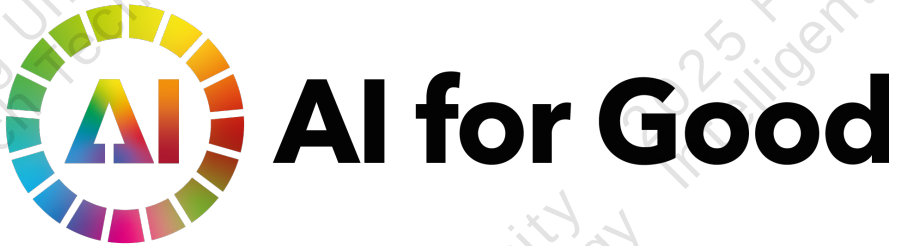


防伪技术：鉴别真伪

音频水印：追踪来源

政策上：积极引导





从语音到声音

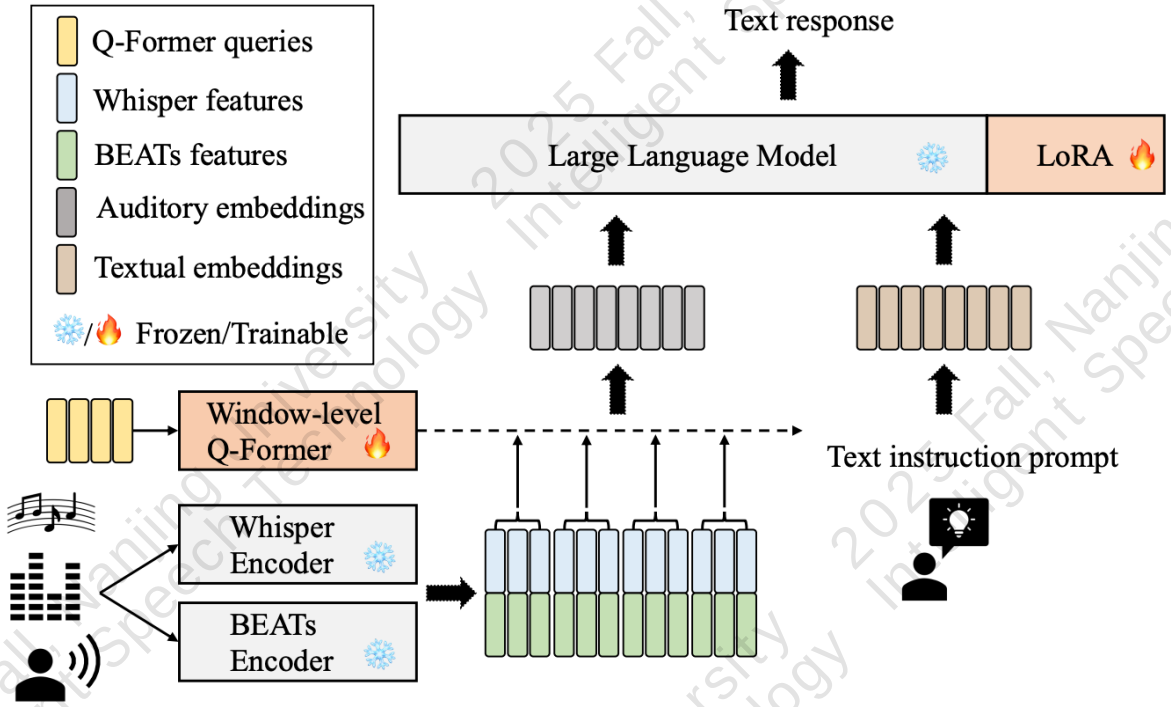
音频生成

拟音师为电影制作场景音效

Meta: MovieGen 音效生成

从语音到声音

音频理解



SALMONN: Towards Generic Hearing Abilities for Large Language Models

音频内容：
Can you guess where I am right now?

请根据背景音详细回答说话者的问题。

模型输出：从背景音判断，说话者似乎身处战区或战斗环境中。背景中能听到枪炮声和爆炸声。说话者正在询问听者能否猜出他们身在何处。

超声波及应用

超声波特性：

1. 频率高（＞20000Hz），波长短
   - ➢ 波长短，可检测细微物体

2. 方向性好，近似直线传播
   - ➢ 超声测距、

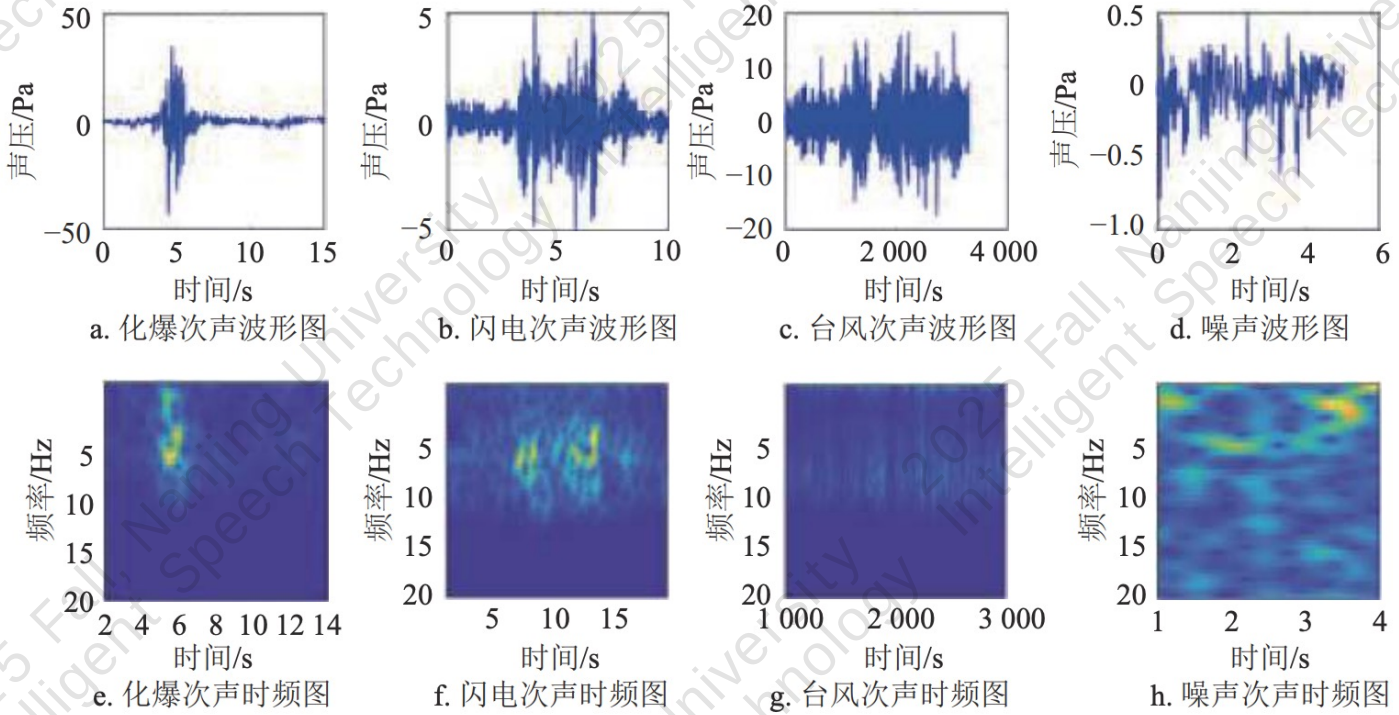
3. 能量集中，用于特定区域作用
   - ➢ 超声清洗、超声焊接
   - ➢ 超声碎石、超声理疗

4. 穿透能力较强，可穿透多种物质
   - ➢ 医学超声检查

超声理疗仪

次声波及应用

1. 频率低（＜20Hz），波长长
   - ➢ 可以绕过障碍物
2. 传播距离远，衰减缓慢
   - ➢ 地震、海啸检测
   - ➢ 军事侦查
3. 穿透力强，可穿透障碍物
   - ➢ 地震次声波穿透地壳
4. 与人体器官共振危害大
   - ➢ 次声武器（人道约束！）



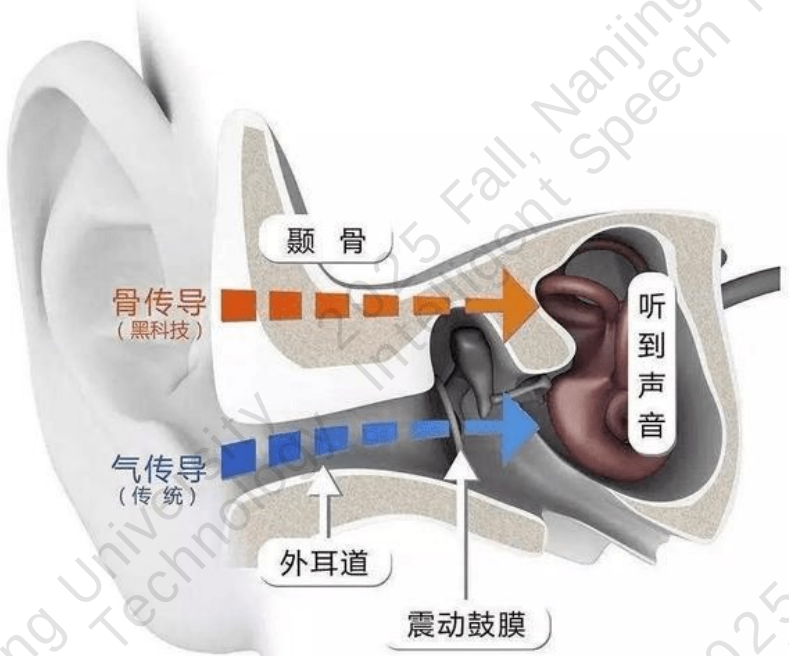
a. 化爆次声波形图　b. 闪电次声波形图　c. 台风次声波形图　d. 噪声波形图

e. 化爆次声时频图　f. 闪电次声时频图　g. 台风次声时频图　h. 噪声次声时频图

大气低频声监测技术是中远程核爆炸探测的技术手段之一，次声监测已被联合国列入全面禁核监测技术

吴浈晖, 赵子天, 陈晓雷, 等. 大气低频声信号识别深度学习方法研究[J]. 电子科技大学学报, 2020,

骨传导语音增强



空气传导 --> 骨传导

利用骨传导技术，通过颅骨将声音直接传递到内耳，从而绕过外耳和中耳的障碍，提高语音的清晰度和可懂度。

无声语音接口

定义

无声语音接口（SSI）是一种创新的人机交互技术，其核心在于**不依赖于声带振动产生可听见的语音**，而是通过捕捉与言语产生过程相关的生理信号，并将其解码为文本或合成语音。

工作原理

- 捕捉**发音器官**（舌头、嘴唇、喉部肌肉）的微弱运动
- 检测**肌肉电活动**（如EMG信号）
- 记录**脑电波**中与语言思维相关的特定模式

与传统语音的区别

传统语音交互

依赖声带振动产生声音，通过空气传播的声波被麦克风捕捉，然后转换为电信号进行处理。

VS

无声语音接口

不产生声音，通过捕捉与言语相关的**生理信号**，直接解码用户意图或语言信息。

无声语音接口

辅助沟通障碍人群

对于因**疾病**（如肌萎缩侧索硬化症ALS、中风、喉切除术等）导致失语或发声困难的患者，SSI提供了一种全新的、**非侵入性或微创性**的沟通途径，极大地改善他们的生活质量和社交参与度。

高噪音环境通信

在嘈杂的**工业生产线、战场、消防现场**或水下等环境中，传统语音通信往往受限。SSI允许用户在不发出声音的情况下进行交流，确保信息传递的清晰性和保密性。

隐私与保密通信

在需要**高度保密**的场合，如军事行动、情报传递或公共场所的私人对话，SSI能够实现"无声"的信息输入，有效避免信息泄露，保障沟通的隐私安全。

提升人机交互效率

在某些场景下，如**公共交通工具上、图书馆或深夜**，用户可能不便发出声音进行语音控制。SSI提供了一种**静默的交互方式**，使得智能设备控制、虚拟现实（VR）/增强现实（AR）交互等更加便捷和私密。

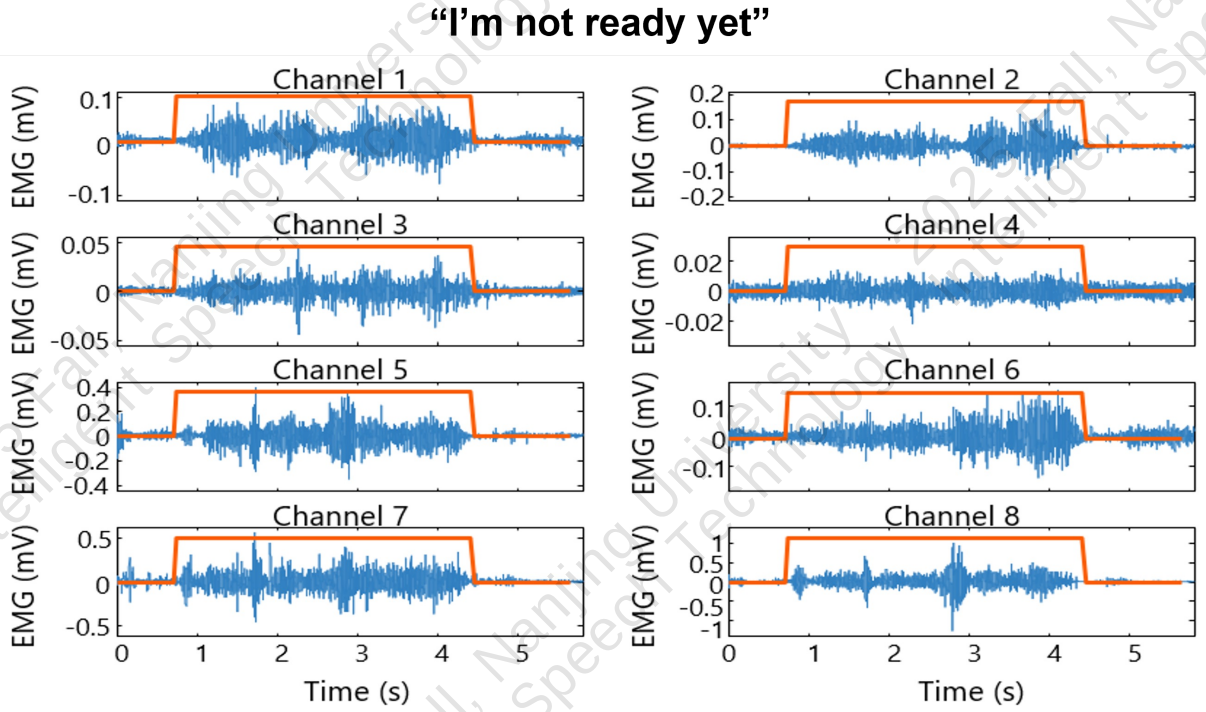
医疗场景：电子喉

替代受损的喉头声带成为新的音源

失去喉部功能的使用者将电子喉的末端放置在颈部颈前侧的最佳发音点上，这里的皮肤、肌肉和骨骼等组织能够将振动传递到咽部。振动通过颈部组织传导至咽部后，引起咽部空气的振动，从而产生声波。

其他形式的“无声语音”

肌电信号（Electromyography, EMG）技术通过检测肌肉活动时产生的微弱电位变化来工作。即使没有发出声音，与发音相关的肌肉仍然会发生微小的收缩，产生可被传感器捕捉到的表面肌电信号（sEMG）。



Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob.

其他形式的“无声语音”

Glasses with AI can read a silent speech

SSI 各类技术路径的对比

特性维度	肌电信号 (EMG)	超声成像 (Ultrasound)	脑机接口 (BCI) – 非侵入式 (EEG)	脑机接口 (BCI) – 侵入式 (ECoG)
设备便携性	较好 ●●●○○ 可集成到腕带、眼镜等可穿戴设备	中等 ●●●○○ 探头和处理单元仍有一定体积	较差 ●●○○○ 需要佩戴头盔或电极帽，舒适度欠佳	极差 ●○○○○ 需要外科手术植入，非便携式
用户适应性	相对容易 ●●●●○ 需用户进行肌肉训练和校准	相对容易 ●●●●○ 需用户保持特定姿势	较难 ●●●○○ 需要长时间训练和专注	极难 ●○○○○ 侵入性手术风险，仅限特定医疗用途
主要优点	• 非侵入性 • 相对便携 • 成本较低	• 非侵入性 • 能捕捉发音器官的物理运动	• 非侵入性 • 理论上可直接读取"思想"	• 解码精度和速度最高 • 最接近自然语音
主要缺点	• 易受肌肉伪影干扰 • 个体差异大 • 连续语音解码难	• 设备体积和舒适度有待提高 • 对操作要求高	• 信噪比低 • 解码精度差 • 易受外界干扰	• 侵入性 • 手术风险 • 伦理争议 • 成本极高