第二届全国社会舆情论坛

智能时代 前进中的语言声学

互联网多媒体内容分析中的音频处理技术

颜永红

中科院语言声学与内容理解重点实验室

2020.9.6

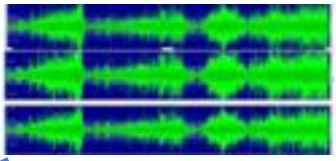
内容提纲

- 互联网多媒体内容分析难点
- 所需语音核心技术简介
- 中科院语言声学与内容理解重点实验室简介

互联网多媒体内容分析难点

- 互联网上音视频内容相比传统信道上内容更为复杂，录制场景、背景噪音多样，常常有多人出现，多种编码格式，伴随着音乐等非语音音频
- 需要每天从上亿规模节目中寻找特定内容/特定人物，其难度无异于大海捞针，特别对识别技术的精度提出了挑战，同时对虚警有更高的要求
- 如何克服信道差异与背景噪音
- 如何解决多人声音同时出现
- 超大规模海量多媒体内容分析处理需要采用更快速的识别技术和策略

音频相关处理技术



音频DNA

语音增强

语种识别

丰富音频检测

音频分析处理相关技术

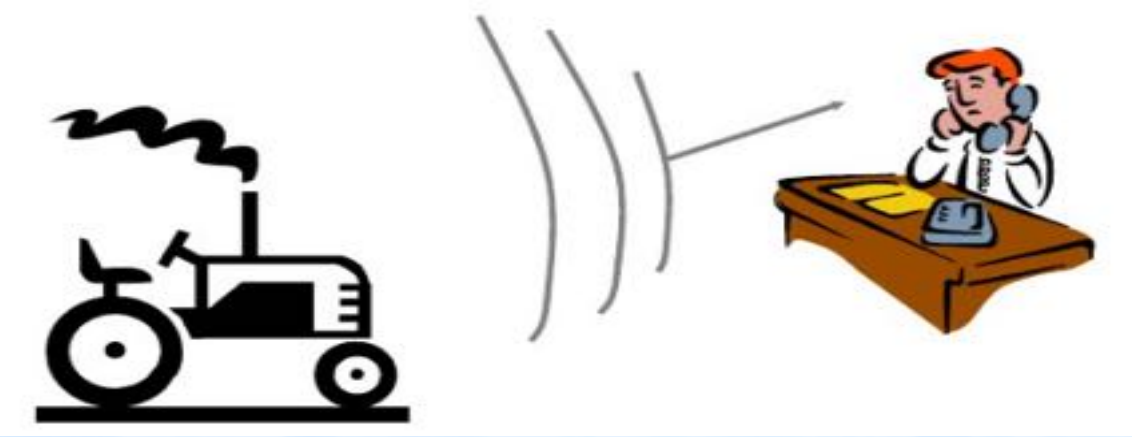
声纹识别

音频内容理解

语音识别

核心技术：语音增强

- 需求：语音在产生和传输过程中，易受各种各样的噪声干扰，严重影响语音识别等技术的性能，如何从含噪语音中提取尽可能纯净的原始语音？
- 定义：语音增强是指当语音信号被各种各样的噪声干扰、甚至淹没后，从噪声背景中提取有用的语音信号，抑制、降低噪声干扰的技术。



核心技术：语音增强

中国科学院声学研究所
Institute of Acoustics, CAS

在现实生活中，很多语音识别的情景并不是干净的声学环境。如右图所示，在一个会议转录场景下，噪声、混响、多说话人始终干扰着识别系统的正常工作。

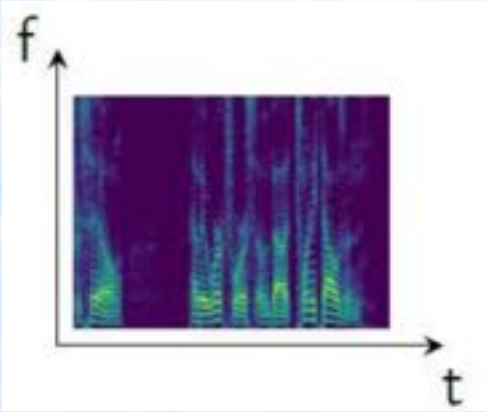
能听清楚在说什么吗

短视频的背景场景更是几乎涵盖了生活中碰到的各种情况！
针对这些问题，在语音增强领域发展了噪声抑制、混响消除、多说话人分离这些领域。而针对语音识别，如何鲁棒地在复杂的真实场景下（会议、车载、家居、户外等）运行是最大的挑战。

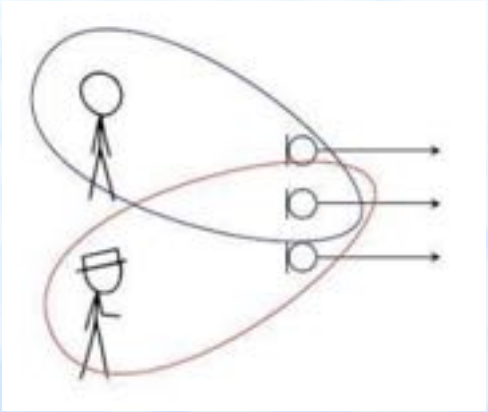
核心技术：语音增强

中国科学院声学研究所
Institute of Acoustics, CAS

- 语音增强侧重于听感质量上的提升，语音识别侧重于识别率的提升
- 两者息息相关，但又有所不同。因为人的大脑对不同形式语音谱的理解大大超过计算机，听感的提升在部分情况下未必能在识别系统上有所体现。
- 如果实现语音增强，以多说话人为例（经典的鸡尾酒会问题）：



不同人有不同的声音特性（如基频、语调等），转化为模式识别问题



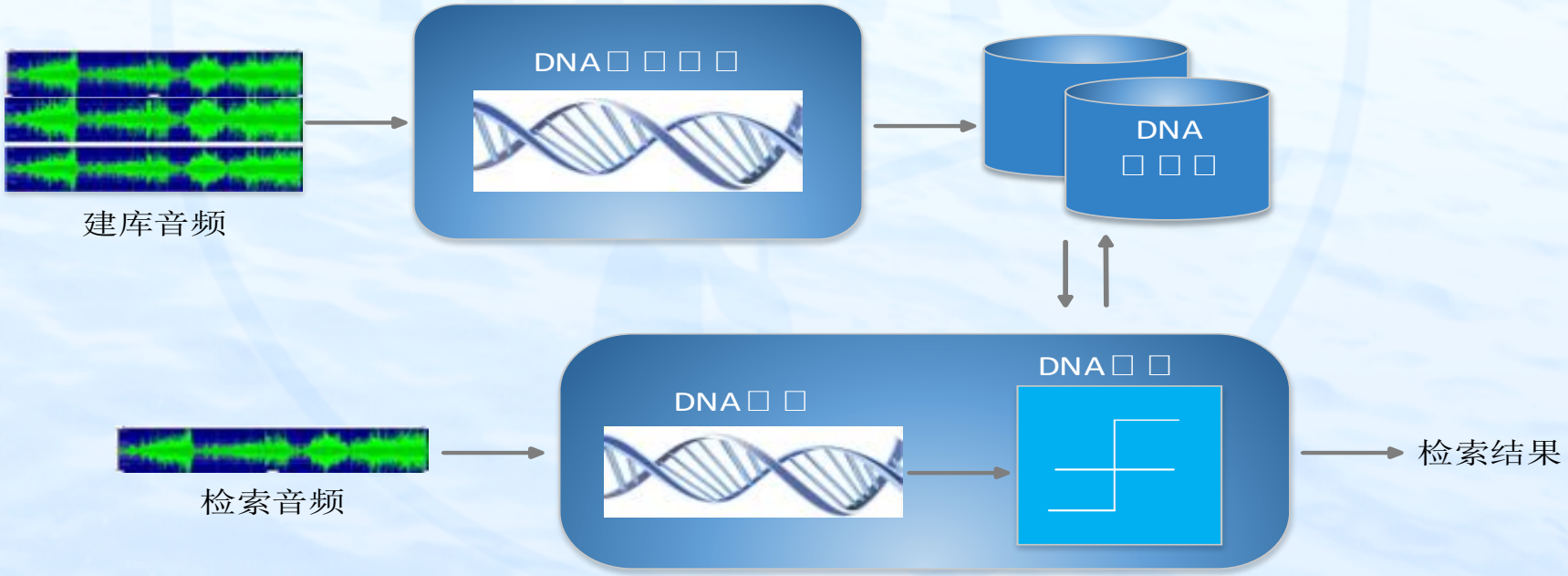
不同人站在不同的位置说话，转化为空间滤波问题（保留指定方向，抑制其他方向）

核心技术：音频DNA

中国科学院声学研究所
Institute of Acoustics, CAS

定义

- 音频DNA，通过分析处理一个音频片段，基于模板匹配技术，从海量数据中检索出包含模板的音频
- 音频DNA检索是一种基于音频内容的检索技术，可以通过直接输入音频片段对音频数据库进行检索，而无需任何文字输入



核心技术：音频DNA

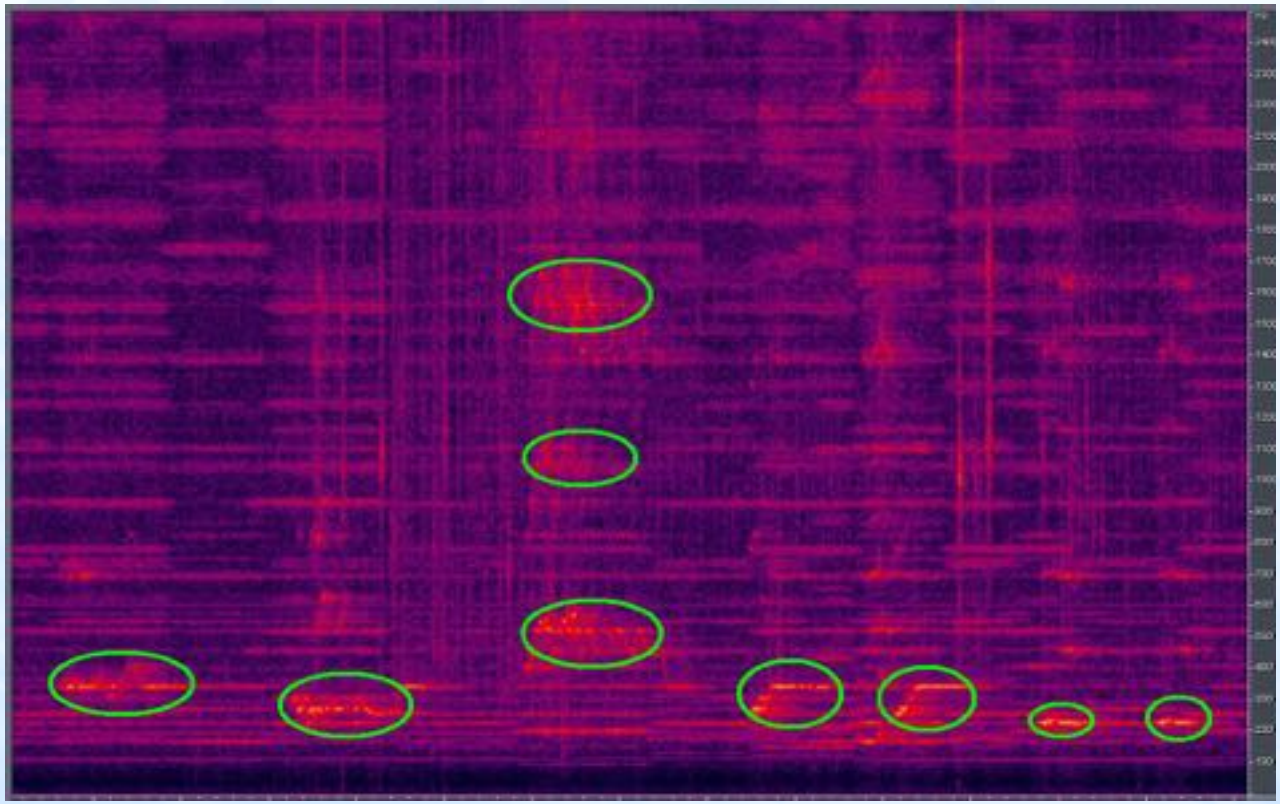
- 基于音频DNA（样例模板匹配）技术，用于快速检索出包含重复片段的语音，进行过滤、去重
- 技术难点
  - 音频质量不同，噪声影响大
  - 检索速度
  - 重复片段位置和时间长度不确定
- 解决方案
  - 优化音频检索特征提取算法，通过筛选谐波能量谱峰值增强对噪声的鲁棒性
  - 优化索引及检索算法

核心技术：音频DNA

中国科学院声学研究所
Institute of Acoustics, CAS

- 依据人耳听觉特性设计滤波器
- 人耳对时间和频率方向变化剧烈点最敏感
- 能量大且差异大的点才能在各种失真中幸存

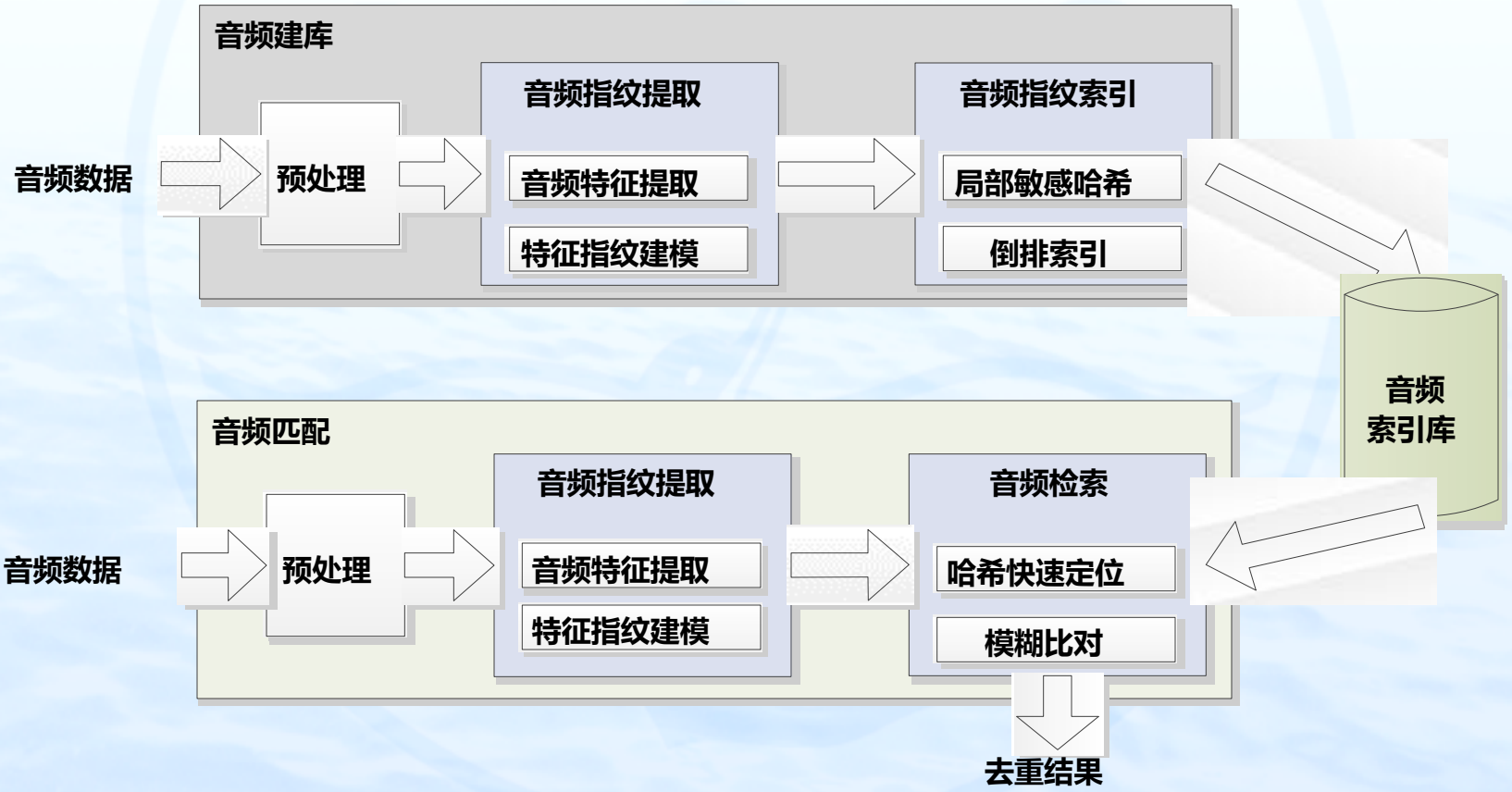
- 优化音频检索特征提取算法
- 加入时序错位差分信息
- 特征掩蔽算法，对不可靠的音频特征进行掩蔽



核心技术：音频DNA

中国科学院声学研究所
Institute of Acoustics, CAS

❑ 音频匹配技术框架图



核心技术：音频DNA

中国科学院声学研究所
Institute of Acoustics, CAS

- 20万模板库（30s）
  - 99%非目标数据
  - 1%目标数据
- 测试片段
  - 随机选取10000个不同长度片段
  - 5种编解码方式，码速率大约为20kbps
- 运行速度
  - 5000倍实时相当于640Mbps

长度	召回率	精度
3秒	96.7%	99.7%
6秒	99.2%	100%
9秒	99.90%	100%
12秒	99.90%	100%

模板数	10万	15万	20万
时长/运行时间（实时率）	6320	6068	6044
内存消耗	2G	3G	4G

音频DNA技术：自动音频模板发现

中国科学院声学研究所
Institute of Acoustics, CAS

- 技术方案
  - 通过音频检索手段找到已有音频数据之间的相似片段
  - 根据相似片段的所属关系进行音频聚类
  - 待测音频与各类别数据比对，分组至具有相同片段的类别或新建类别
- 应用场景
  - 冗余音频数据消除
  - 音频代表性样本推荐功能
  - 白名单

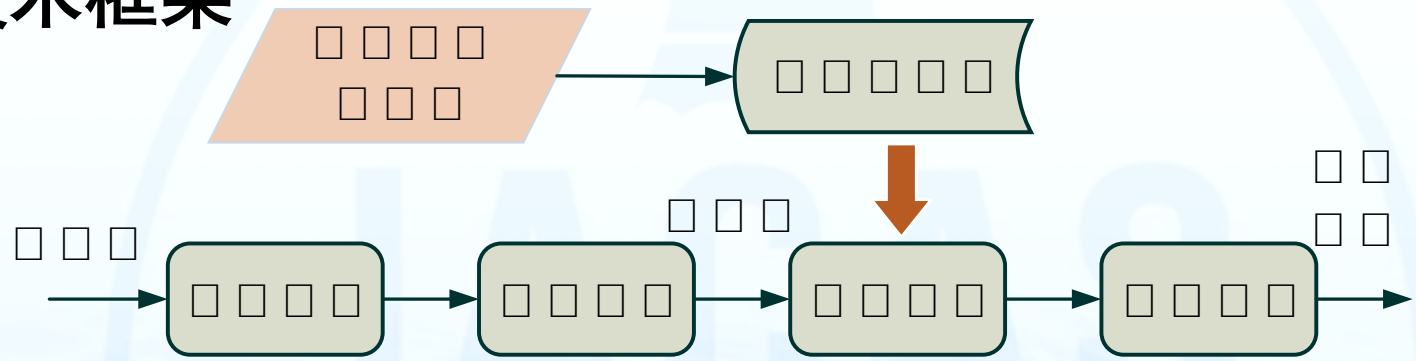
核心技术：语种识别技术

- 定义
  - 它是通过分析处理一个语音片段以判断其属于某个语言种类的过程，其本质是语音识别的一个方面
  - 简而言之，就是识别出“用什么语言说的”
- 用途
  - 克服人工监测的局限性：由于人的精力是有限的，不可能同时掌握很多语种
  - 多语种人工接听，跨语种语音识别，信息安全
  - 可以从海量的互联网音视频实时流中发现目标语种节目并自动分类标记

核心技术：语种识别技术

中国科学院声学研究所
Institute of Acoustics, CAS

技术框架



关键技术

- 特征提取
  - MFCC，LPCC，MFPLP
  - 短时差分 delta变换
  - 长时移动差分变换(SDC)
- 目标语种建模
  - 高斯混合模型（GMM）
  - 支持向量机（SVM）
  - N元文法建模
- 概率计算
  - 似然概率
  - 分类面距离度量
  - N元文法似然概率计算
- 判决规则
  - 假设检验
  - 全局门限
  - 判决风险

核心技术：语种识别技术

中国科学院声学研究所
Institute of Acoustics, CAS

- 复杂音频的挑战：实网数据（特别是互联网音视频）包括大量音乐、背景音乐、各种噪声、色情等非有效语音，如何提取有效部分进行识别
- 信道差异的影响：各类音频节目产生和传输的方式不同，如何消除实际场景数据和训练数据之间的信道不匹配
- 易混淆语种/方言的区分：互联网海量数据流中小语种占比极小、易混淆语种区分性差，如何才能准确快速的检测小语种？

核心技术：语种识别技术

中国科学院声学研究所
Institute of Acoustics, CAS

针对难题	采用关键技术点	实现
互联网音频复杂多样，如何提取有效语音	利用声学特性+垃圾反模型相结合的策略	提升垃圾过滤和有效语音检测能力
针对难题	采用关键技术点	实现
对复杂音频，如何消除信道和噪声影响	利用多种特征补偿技术，消除信道差异	得到鲁棒的语种特征，提升识别性能
针对难题	提出	实现
易混淆语种区分性差	基于语支鉴别性的语种识别方法	提升了易混淆语种的鉴别性

核心技术：语种识别技术

- 支持语种涵盖国内外大语种：汉语、英语、法语、俄语、德语、日语、韩语、维语、藏语、粤语和闽南语等二十几种。

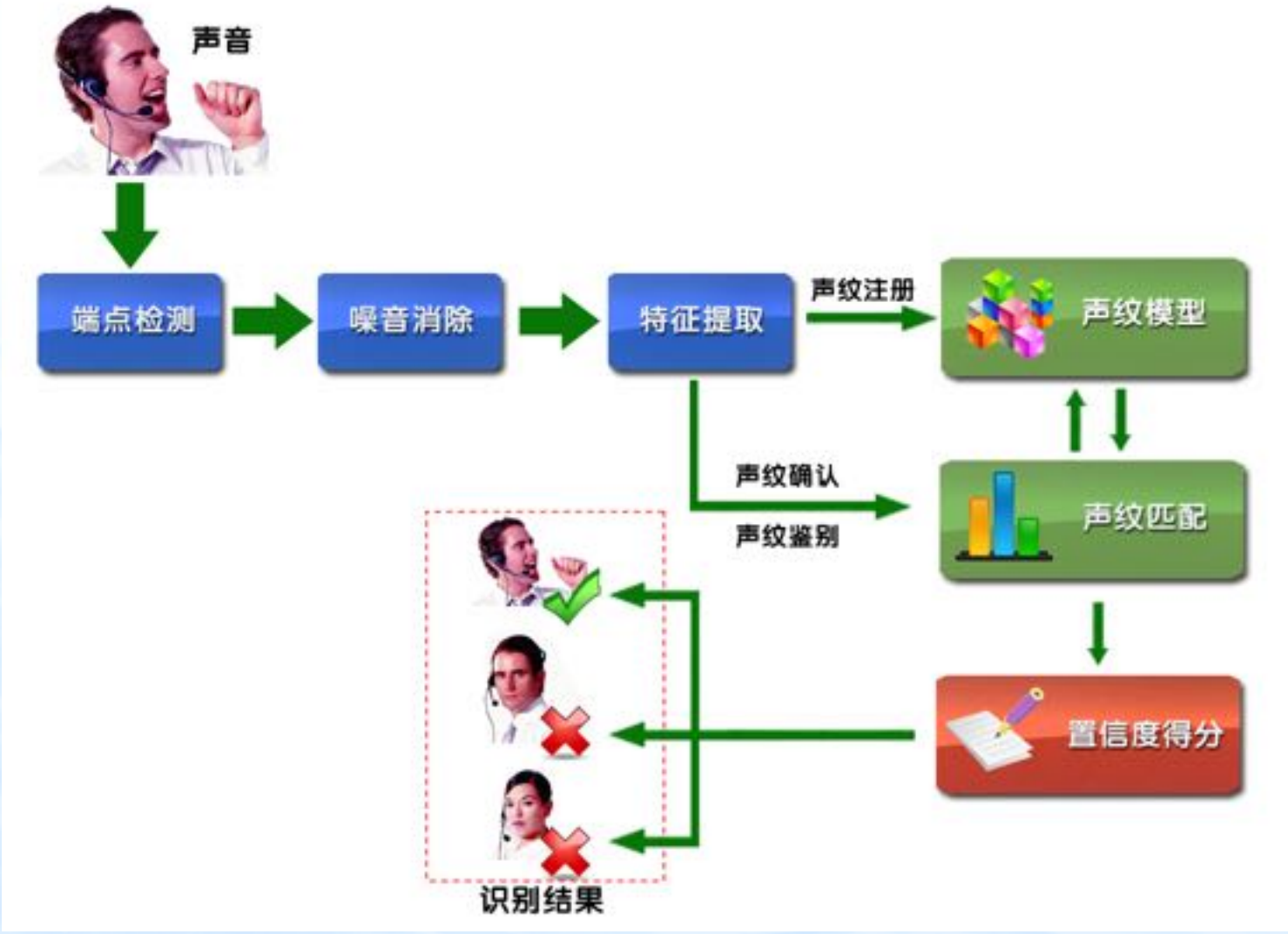
- 支持国内重点方言：四川、北京、厦门、上海等地区方言；
- 限定目标语种，语种识别正确率已达到95%以上；
- 目前语种识别技术面临的主要难点包括：短时语音、集外拒识 和 复杂声学环境，也是当前技术研究的重点。

核心技术：说话人识别

- 定义
  - 说话人识别是一种根据语音中反映的代表说话人生理和行为特征的语音参数，来自动识别说话人身份的技术。
  - 识别当前输入语音的话者身份，简而言之就是识别出“谁说的”
  - 说话人识别包括：注册、测试

- 说话人识别技术也是生物识别技术的一种，通过生理或者行为特征对人的身份进行识别
  - 每个人都应该拥有该特征
  - 该特征对每个人都具有明显的区分性
  - 该特征在一定的时期内固定不变
  - 该特征易于采集

核心技术：说话人识别



核心技术：说话人识别

中国科学院声学研究所
Institute of Acoustics, CAS

- 与其他生物识别技术（如脸型、掌形、虹膜识别等）相比较：
- 提升用户体验
  - 不涉及隐私，用户无心理障碍，用户接受度高
  - 在自然对话中即可实现声纹识别

- 远程控制便捷
  - **非接触式识别**，唯一可用于远程控制的生物识别技术
  - 安全可靠、方便便捷

核心技术：说话人识别

- **跨信道问题**：
  - 信道多样，复杂性导致的准确率及召回率偏低
- **多说话人问题**：
  - 当识别语音流中存在有多个说话人时，会导致说话人识别，准确率和召回偏低（说话人自动分段聚类）
- **目标说话人注册语音的影响**：
  - 当目标说话人注册语音偏少，且注册语音中存在频谱缺失时，靠人工基本无法一一定位，但是这种注册语音容易导致声纹特征不能代表目标人的真实特征，会造成识别准确率偏低

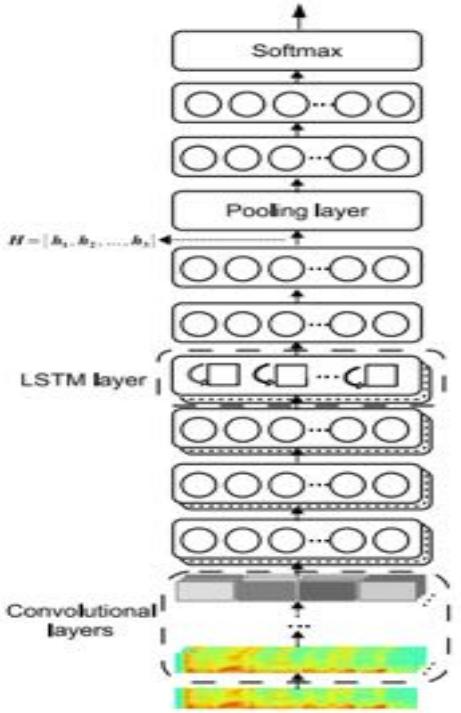
核心技术：说话人识别

中国科学院声学研究所
Institute of Acoustics, CAS

- 时序累积建模
  - 基于注意力机制的时频域信息融合
  - 特征提取、时序累积、鉴别建模分布学习
    - 提高模型鲁棒性和长时信息学习能力

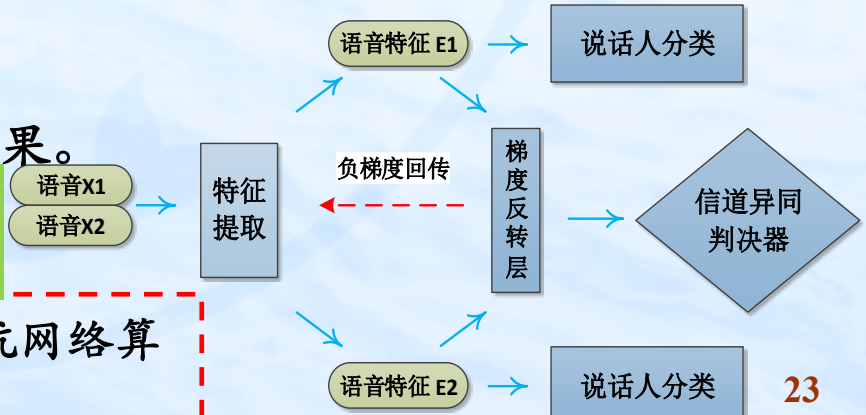
X. Miao, etc, “*A new time-frequency attention mechanism for TDNN and CNN-LSTM-TDNN, with application to language identification*”, Interspeech 2019.

弥补TDNN上下文时序累积较短的问题



- 跨信道补偿
  - 孪生网络成对学习
  - 信道异同判决器
    - 克服训练数据不平衡问题
    - 提升在未知信道场景下的识别效果。

Chen Zhigao, et al, *Cross-Domain Speaker Recognition Using Domain Adversarial Siamese Network with a Domain Discriminator*, Electronics Letters, 2020(SCI)



业界目前最好的跨信道补偿算法，相比一般对抗网络算法性能相对提升20%。

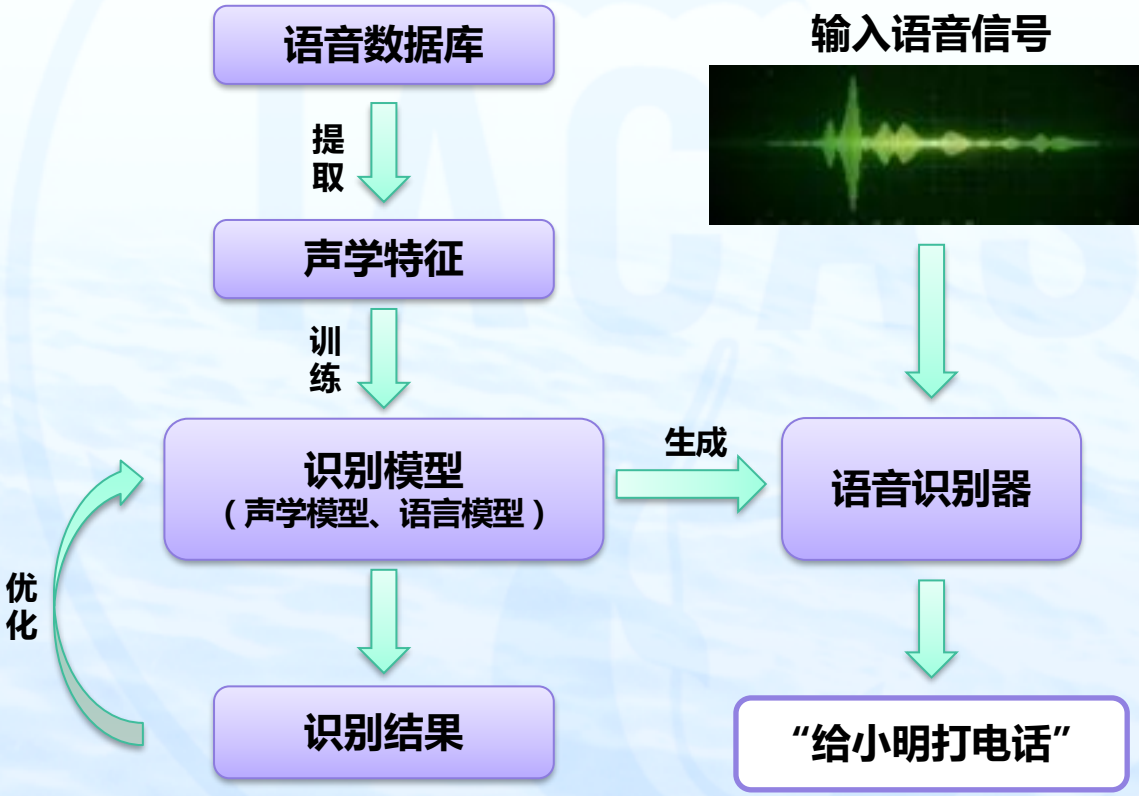
核心技术：说话人识别

- 短时文本相关声纹识别等错误率控制在1%以内，已经大规模实用
- 长时文本无关声纹识别等错误率控制在2%以内，在海量数据检索领域已经实用，随着目标人样本数据丰富，目标人识别精度会逐步提高
- 复杂声学环境下（如跨信道情况），声纹识别技术依然面临挑战，也是目前的研究重点

核心技术：语音识别技术

- 需求：如何在海量音视频中实现基于内容的分析？
- 定义：语音识别（ASR）通过对语音信号进行处理转成文字内容。
- 互联网中包含多语言数据：开展多语言语音识别统一框架研究，快速研制针对小语种（如维语）的语音识别系统。
- 复杂声学环境/自然口语对话：研制关键词检测系统，通过检测敏感关键词实现内容分析。

核心技术：语音识别技术


语音数据库
提取
声学特征
训练
识别模型
（声学模型、语言模型）
优化
识别结果
生成
输入语音信号
语音识别器
“给小明打电话”

核心技术：语音识别技术

中国科学院声学研究所
Institute of Acoustics, CAS

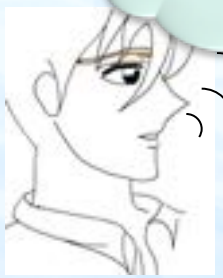
$$= \sum_{k=0}^{i} c_k \frac{1}{\sqrt{2\pi}\sigma} e^{-(o_t-\mu)^2/2\sigma^2}$$

$$= \prod_{k=0}^{M-1} P(w_k | W_{k-N+1} \cdots W_{k-1})$$

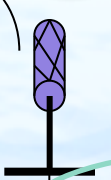
识别过程是基 音词典、
个使后验概率 化的词串作为识别结果

$$F = \quad (u)) = \arg\max_i \frac{P_\lambda(O(u) \mid w_i(u))\ P(w_i(u))}{P(O(u))}$$

建模单元的选择
概率模型的选择
模型的训练方法



用户话音



前端处理

声学模型

语言模型

解码器

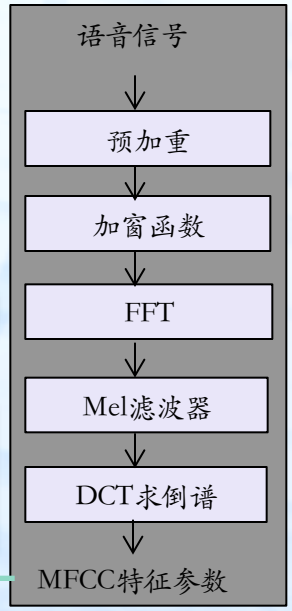
识别结果

发音词典

中国：
zh ong1 g uo2
人民：
r en2 m in2

搜索空间的构建
快速算法的研究

特征矢量：$O(u) = \{O_1, O_2, \cdots O_T\}$



核心技术：语音识别技术

中国科学院声学研究所
Institute of Acoustics, CAS

用户语音

前端处理

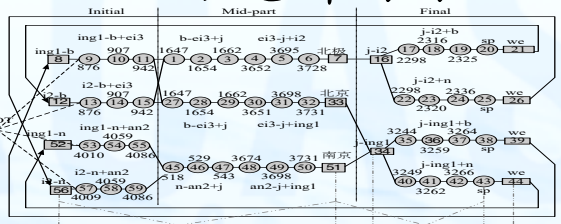
解码器

识别结果

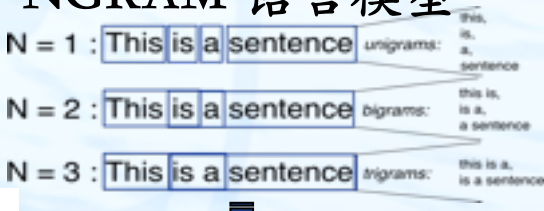
声学模型

语言模型

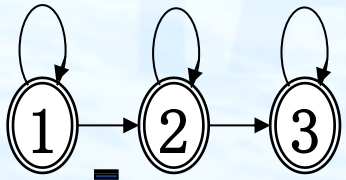
动态解码网络



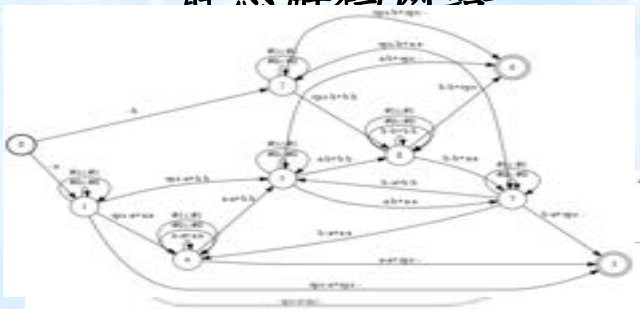
NGRAM 语言模型



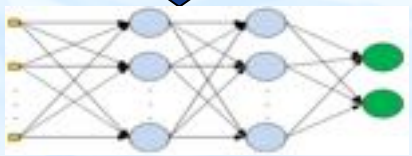
HMM
GMM



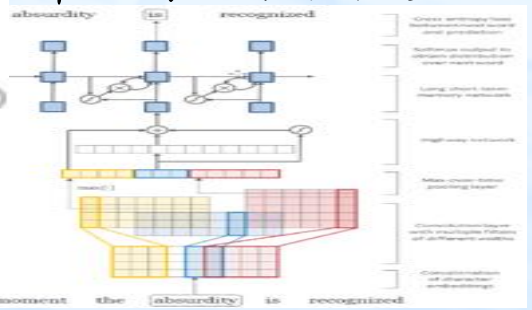
静态解码网络



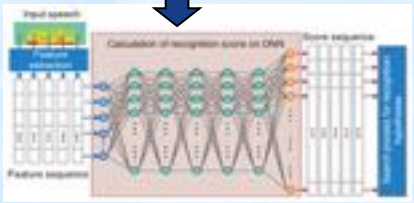
神经网络



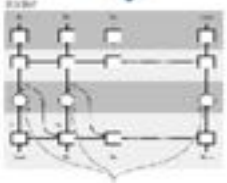
神经网络语言模型



深度神经网络



端到端模型



核心技术：语音识别技术

中国科学院声学研究所
Institute of Acoustics, CAS

- 语音搜索（朗读风格）准确率已达到95%以上，可以实用，未来搜索引擎中，通过语音进行搜索的比例会逐步提高。
- 一般安静环境下，自然口语对话识别准确率已在85%以上，有待进一步提高。
- 复杂声学环境下（如互联网音视频），自然口语对话识别准确率下降比较厉害，这时比较实用的技术是采用关键词检索，通过特定关键词的检测，进行语音内容分析。

核心技术：关键词识别

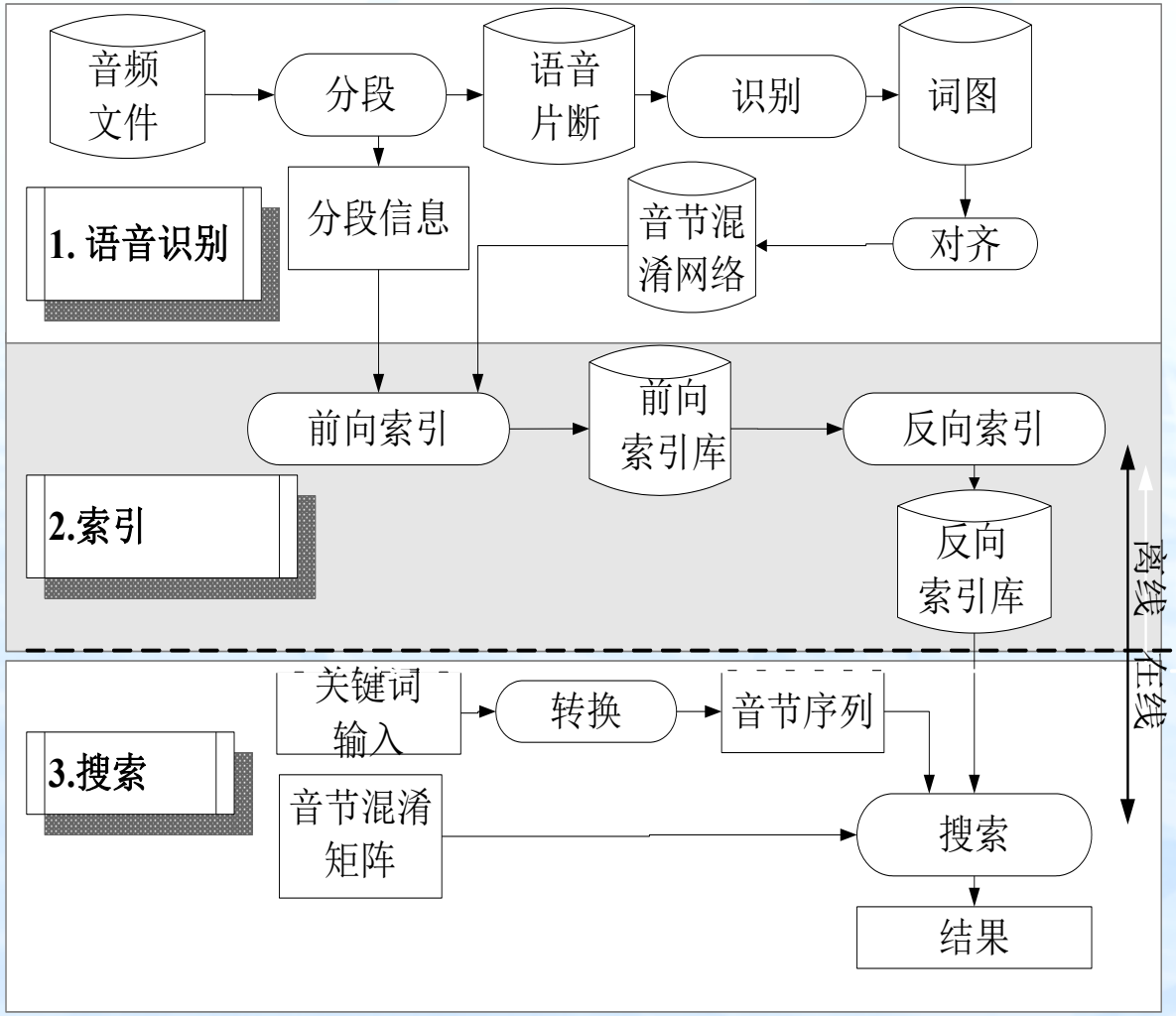
中国科学院声学研究所
Institute of Acoustics, CAS

- 定义：识别当前输入语音中包含用户关心的敏感词语，并定位他们出现位置的技术，同识别不同，其实是一个检索

- 用途
  - 通过设置目标关键词，系统够可以从海量音频内容中发现目标关键词并给出具体位置信息
  - 可以实现对海量音频内容的快速检索和定位

核心技术：关键词识别

中国科学院声学研究所
Institute of Acoustics, CAS

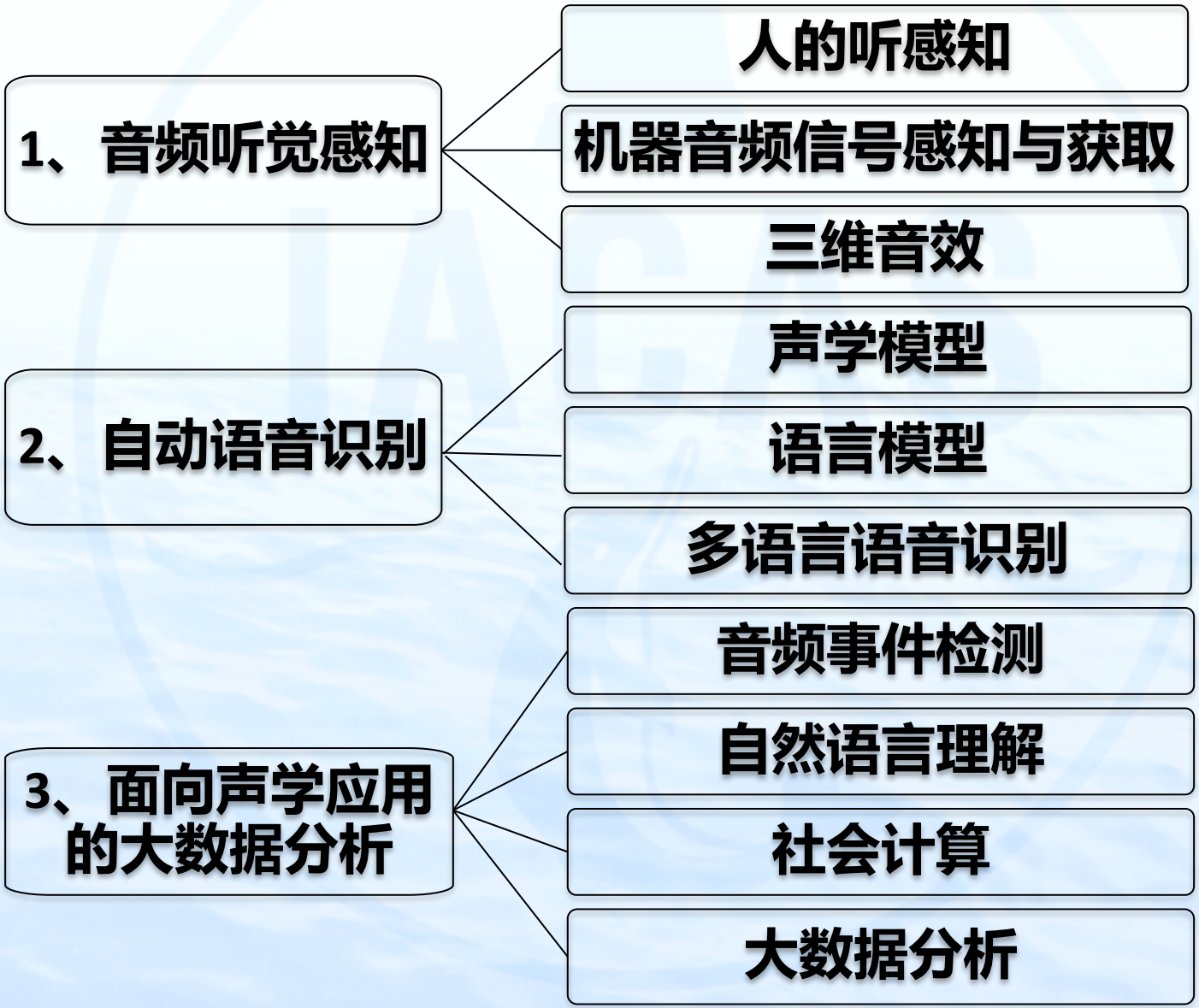


如图，整个检索系统包括三个阶段，分别是：语音识别，索引建立和搜索。在第一个阶段，针对输入的语音流(语音库)，进行语音分段，然后进行识别，得到音节混淆网络数据结构；第二个阶段，利用音节混淆网络的单个弧路径建立逆索引结构；第三个阶段，应用音节混淆矩阵来计算相关分值，利用快速模糊搜索算法来提取关键词候选。

最后的搜索阶段才需要目标关键词作为输入！

研究方向

中国科学院声学研究所
Institute of Acoustics, CAS



团队业绩、学术水平

- 建设了本学科世界一流的研究设施（通用服务器200余台、500余个GPU，数据存储能力2.5P ）
- 国内863公开评测两次第一，美国国防部语音评测三次世界第一，国际音乐检索组织评测六次世界第一；为百度、腾讯、阿里巴巴、华为、联想等提供了语音核心技术
- 累计获得授权发明专利101项、正在受理中专利40余项，软件著作权登记50余项，发表论文500余篇（其中SCI120余篇），获得省级科技进步一等奖3项、中国科学院杰出科技成就奖1项、国家科技进步二等奖一项

- 团队承接的具有代表性本学科国家项目
  - 国家杰出青年基金，语音领域第一个
  - 国家自然科学基金重大项目“多语言言语识别基础理论与建模方法”， 语音领域第一个
  - 中科院战略科技先导专项——媒体大数据项目牵头单位（2.3亿）

相关技术评测成绩

中国科学院声学研究所
Institute of Acoustics, CAS

全国首届网络舆情（音视频）分析技术邀请赛共收到来自全国高校、科研院所、互联网企业和单位共计141支队伍报名，经过赛事组织委员会审核，正式邀请清华大学、北京大学、复旦大学、中科院声学所、公安部三所、商汤科技、富士通等32支队伍参加8项赛事的角逐。

此次比赛重点面向互联网音视频分析的应用背景，设置音频比对、说话人识别、语音关键词检测、拷贝视频检测、特定视频识别、视频文本关键词检测、人脸识别和流媒体检测等项目。比赛评测

基于深度学习技术搭建的语音/音频识别系统取得的成绩：

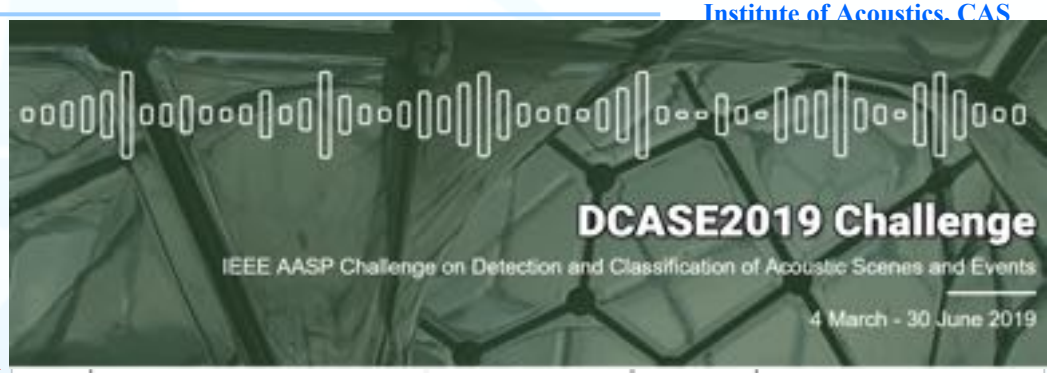
在2016年11月全国网络舆情（音视频）分析技术评测中，取得了所有音频类比赛的第一名

音频场景国际比赛第一（DCASE2019）

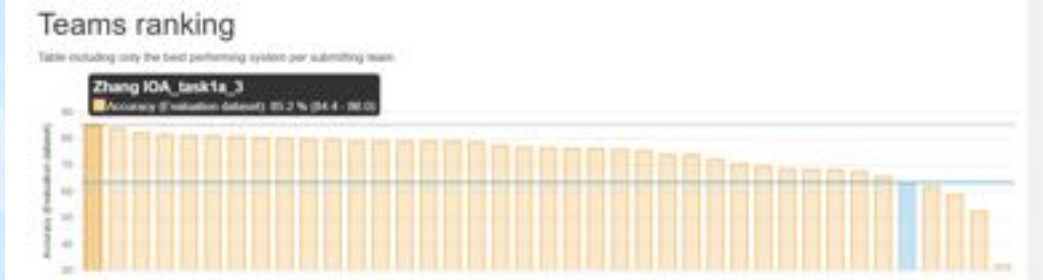
中国科学院声学研究所
Institute of Acoustics, CAS

2019年6月，声学所语音团队在音频场景分类DCASE2019比赛中夺冠。在参加的Task1A任务官方最终排名中，准确率达到85.2%，领先第2名1.4%的绝对准确率，远超人类的分辨能力（75%）。

DCASE 是IEEE 声学信号处理技术委员会（AASP）组织，伦敦大学等2013年发起，是目前世界范围内最大规模的音频分类与检测比赛，目的是识别出录制音频的特定场景，如地铁、公园、机场等，从而使穿戴式设备、智能机器人感知周围的环境信息并做出反应。



Rank	Submission code	Submission name	Technical Report	Accuracy with 95% confidence interval (Evaluation dataset)
1	Zhang_IOA_task1a_3	ZhangIOA3		85.2 % (84.4 - 86.0)
2	Koutini_CPJKU_task1a_4	variants2		83.8 % (82.9 - 84.6)
3	Seo_LGE_task1a_4	LGE_ROBOTICS		82.5 % (81.7 - 83.4)
4	Yang_UESTC_task1a_2	6models-4folds_rf		81.6 % (80.7 - 82.5)
5	Huang_IL_task1a_3	DLensemble		81.3 % (80.4 - 82.2)
6	Jung_UOS_task1a_4	S_KD_4		81.2 % (80.3 - 82.1)
7	Wang_NWPU_task1a_1	Mou_task1A1		80.6 % (79.7 - 81.5)
8	McDonnell_USA_task1a_2	UniSA_1a2		80.5 % (79.6 - 81.4)
9	Wu_CUHK_task1a_1	Wu_CUHK_1		80.1 % (79.1 - 81.0)
10	Liu_SCUT_task1a_2	HS		79.9 % (79.0 - 80.8)



近期亮点工作：声学建模核心算法

中国科学院声学研究所
Institute of Acoustics, CAS

- Per-frame dropout（帧级别丢弃算法）

通过帧级别的随机扰动阻断固定历史轨迹的形成，解决了序列化训练下，LSTM对不同长度语音输入的不鲁棒性问题。

该成果已被被主流语音识别开源软件kaldi采纳，作为推荐的LSTM正则算法使用。

- SOC（半正交限定技术）

传统的低秩分解技术在压缩模型参数量的同时会导致特征值空间分布极化现象，SOC技术可以降低神经网络的优化难度，增强神经网络的收敛效果。

该成果已被被主流语音识别开源软件kaldi采纳，基于SOC技术的TDNN-F模型被多支参加Chime-5（语音界著名的多通道语音识别比赛）的队伍采用。

- OPGRU（带输出门的低维映射门控递归神经网络单元）

目前语音识别主流的神经网络结构LSTM存在着复杂度高，训练难度大，在特定情况下泛化型差的问题。OPGRU是针对如上问题，在不损失识别精度的情况下的一种替换方案。

TDNN-OPGRU作为一种新型的递归神经网络结构被Kaldi采纳并推广。

近期亮点工作：语音信号处理与识别工具包

中国科学院声学研究所
Institute of Acoustics, CAS

- 研发了自主可控的ThinkITSpeechPlatform
  - 研发新一代自主可控语音信息处理平台，摆脱对国外核心技术平台的依赖（2000-2012年英国HTK、2012-2019年美国Kaldi、2018年至今日本ESPNet）
  - 自主可控语音信号处理平台的研发与设计自主可控，支持高效训练与推理部署

技术平台	训练时长*	推理速度（RTF）	识别精度（Cor）***
Kaldi**	~50天	0.3	81.3
ESPNet	~11天	3.5	None
ThinkITSP	~3天	0.5	82.4（无语言模型）

语音识别系统的训练速度提速16倍

近期亮点工作:智能语音能力云平台

中国科学院声学研究所
Institute of Acoustics, CAS

- **协议接口**

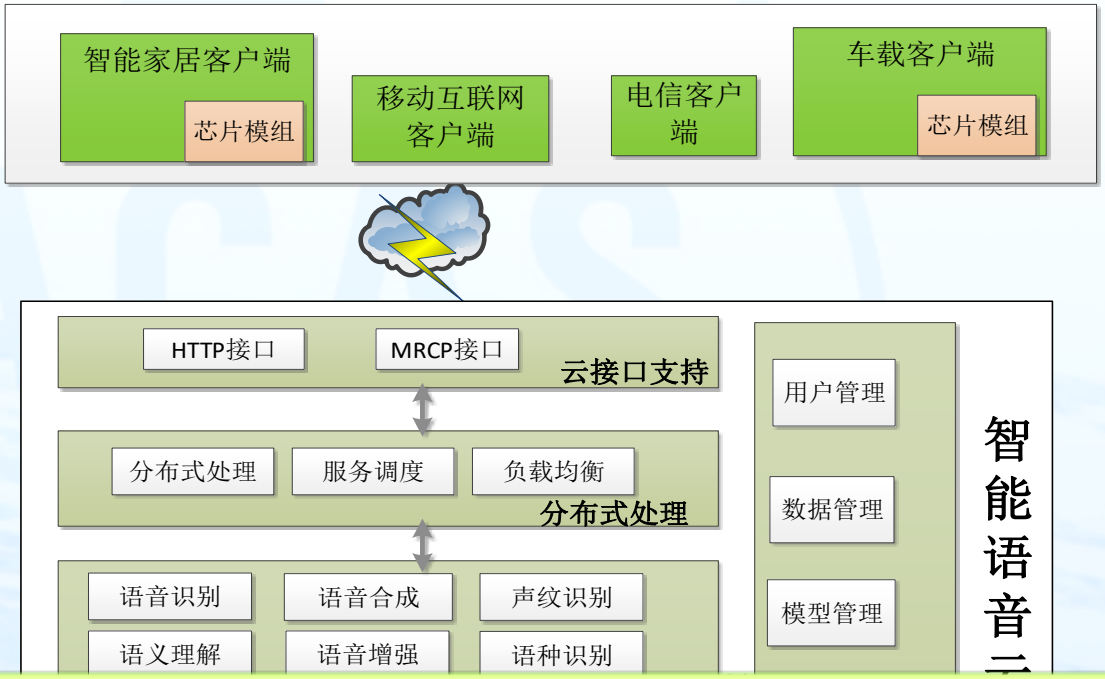
  语音云系统通过HTTP接口提供互联网WEB服务，通过MRCP接口提供电信网服务

- **分布式处理**

  支持负载均衡和分布式扩展，支持高可用和容灾热备

- **系统运维管理**

  提供用户管理、数据管理、模型管理、日志管理、运维监控和统计分析等功能



智能语音能力平台已在中移在线(23个省)、中国电信（20个省）、联通智网、苏宁、平安、广发银行、阳光保险等公司大规模上线应用，支撑用户数超过5亿人

谢谢！