物联网安全
Internet of Things Security

专题1：语音安全

冀晓宇
浙江大学

目录

- 物联网语音基础知识
- 物联网语音安全定义
- 物联网语音安全
  - 语音信号感知安全
  - 语音内容识别安全
  - 语音声纹识别安全
  - 语音意图理解安全

语音基本概念

- **什么是语音**
  - 在语言学中，语音可以被认为是用来**表示语言的声音符号**，是人的发音器官所发出来的具有一定意义的声音
  - 语音是一种**有意义的声音信号**，它在自然界中以声波形式存在，可**转换为电信号**，并通过采样等方式**以数字信息**的形式保存

[1] https://zh.wikipedia.org/wiki/%E8%AA%9E%E9%9F%B3

语音基本概念

- **语音处理：**
  - 又称**语音信号处理**，是研究语音发声过程、语音信号的统计特性、语音的自动识别、机器合成以及语音感知等各种处理技术的总称
- **语音处理的目的：**
  - 从语音信号中计算出一系列参数，高效**传输**或**存储**语音编码，或通过分析计算实现特定用途，如**语音合成**，**说话人辨识**等

Q：语音如何传输?

语音信号处理的发展

- **阶段一：萌芽阶段**
  - 1876年贝尔发明电话
    - 首次用声电、电声转换技术实现了远距离语音信号传输
  - 1939年Bell实验室Homer Dudley发明第一个声码器（vocoder）
    - 在发送端，从话音中提取**参数特征**；到接收端，根据**参数重新合成**，但是合成话语的自然度较差
    - 首次提出了话音模型的思想（参数模型）
  - 1947年Bell实验室发明了**语谱图仪**，自动语言识别ASR (Automatic Speech Recognition)开始，刚开始时是人工分析完成语音识别

语音信号处理的发展

- **阶段二：语音合成、识别时代**
  - 1948年美国Haskins实验室研制成功了“语图回放机”，将语谱图转化为语音，诞生了**共振峰语音合成方法**
  - 1952年BELL实验室Davis等首次研制成功识别十个英语数字的实验装置（根据第一、二共振峰位置特征）
  - 1956年Duddley等人将语音分割成**元音、辅音**等，改进识别装置
  - 1956年Olson等采用8个带通滤波器提取频谱参量作特征，研制成一台简单的声控打字机
  - 1960年Fant发表的开创性工作“语音产生的声学理论”

语音信号处理的发展

- **阶段三：语音处理腾飞时代**
  - 1960、1970年代数字信号处理算法的突破
  - 1965年快速傅里叶（FFT）算法
    - 首次用声电、电声转换技术实现了远距离语音传输
    - 1970年代初动态时间规整（DTW），隐马尔科夫模型（HMM）
    - 1970年代初美国DARPA启动语音理解系统研究计划
  - 此后，人工智能、模式识别、神经网络、机器学习等新技术手段开始进入了语音处理领域

语音在物联网时代的发展

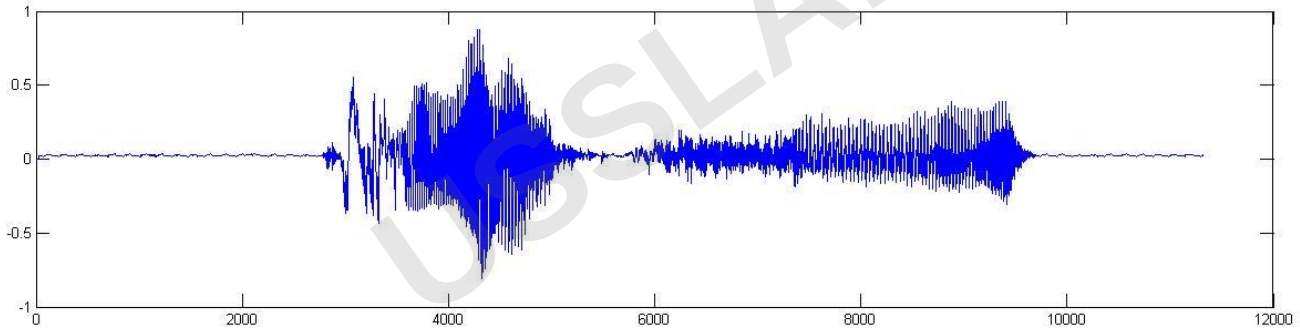
- **技术进展：语音识别**
  - 解决语音识别初期面临的**三大问题**： 依赖说话人、连续/断续发音、词汇量大小
  - 在安静背景下取得了较高识别性能
  - 实现了不同语言环境下不同词汇量的识别

- **语音识别技术取得进展主要原因**
  - 使用统计学习技术：基于隐马尔科夫模型(HMM)
  - 海量语音和文本数据库（语言模型）
  - 高速并行计算能力（云计算）

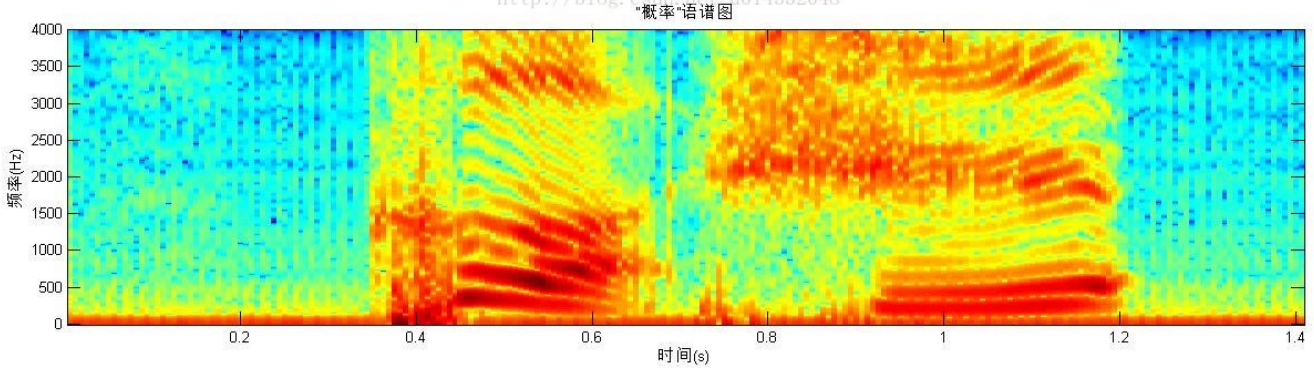
语谱图

- **语谱图：**语音信号的傅里叶分析的显示图形，英文为sonogram或者spectrogram
- X轴时间、Y轴频率、Z轴能量

时域语音信号

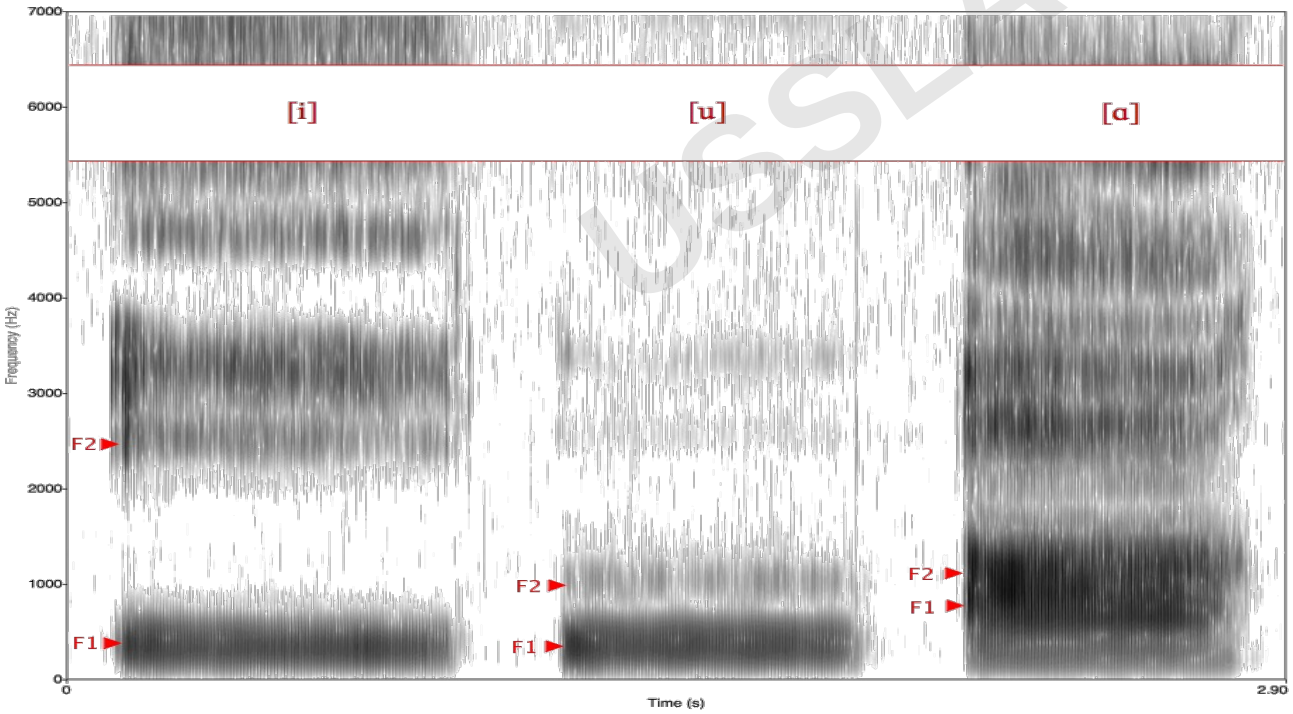


语音信号语谱图



共振峰

- **共振峰（formant）：**在语音科学及语音学中，描述的是人类声道中的声学共振情形，**是区分元音的关键**
- 频率最低的共振峰频率称为 f_1，第二低的是 f_2
- 为什么会有共振峰?



美式英语元音[i, u, ɑ]的声谱图，图中显示了共振峰f_1 和 f_2

元音平均共振峰[7]

元音 (IPA)	共振峰 f_1	共振峰 f_2
i	240 Hz	2400 Hz
y	235 Hz	2100 Hz
e	390 Hz	2300 Hz
ø	370 Hz	1900 Hz
ɛ	610 Hz	1900 Hz
œ	585 Hz	1710 Hz
a	850 Hz	1610 Hz
æ	820 Hz	1530 Hz
ɑ	750 Hz	940 Hz
ɒ	700 Hz	760 Hz
ʌ	600 Hz	1170 Hz
ɔ	500 Hz	700 Hz
ɤ	460 Hz	1310 Hz
o	360 Hz	640 Hz
ɯ	300 Hz	1390 Hz
u	250 Hz	595 Hz

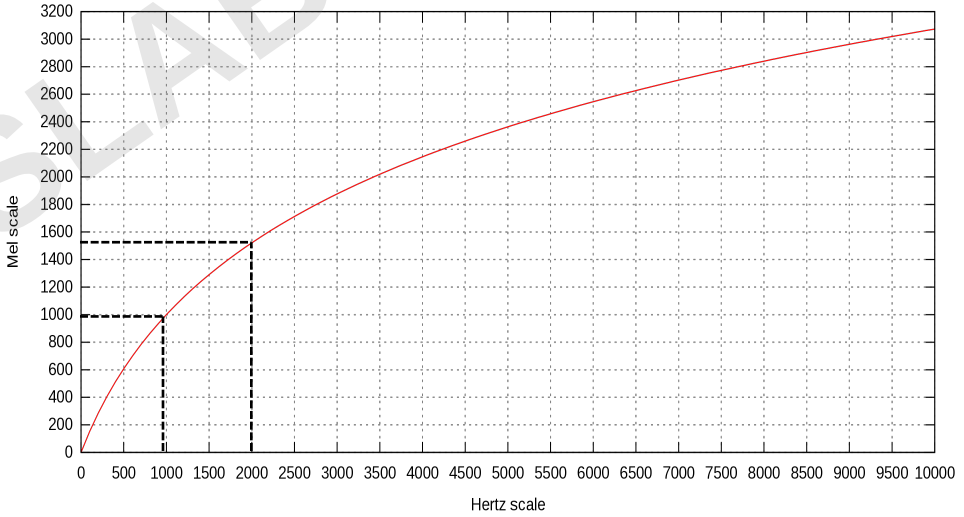
人耳非线性效应

- **人耳非线性**：人耳特性与其他换能器一样，带有非线性的特点，一般用耳蜗的非线性反应解释
- 例如，如果人耳已经适应1000Hz的音调，此时把音调频率提高到2000Hz，我们的耳朵只能觉察到频率提高了一点点，而非提高一倍
- 此外，人耳听到的声音是“合音”，包括多个频率的**差音**、**和音**
  - 例如：当两个纯音同时以400Hz和500Hz同时发出时，仔细听起来还有频率为其差值（100）及其谐频的差值（200,300）的音
  - 和音的频率则是原来两纯音频率之和，在这一例子中就是900Hz
- 在适当频率和强度关系下，一个音可以抑制另一个音的响应（感觉），即**声学掩蔽效应**

Demo: http://szhorvat.net/pelican/combination-tones.html

梅尔倒谱系数（MFCC）

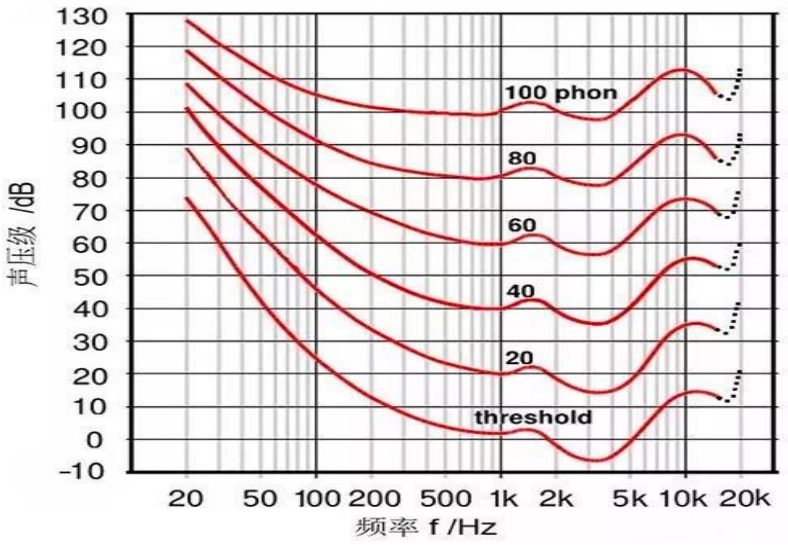
- **梅尔倒谱**：由于人耳对于频率感知非线性，梅尔刻度与线性频率刻度赫兹(Hz)之间可以进行数学换算，使在mel域内人耳听觉为线性：
- 计算公式：Mel=2595*log10(1+f/700)
- 特点：
  - 低频区分辨率高（0-1000Hz）
  - 高频区分辨率低（>1000Hz）



- **梅尔倒谱系数**：mel-frequency cepstral coefficient(MFCC)
  - 倒谱：指对声音的频谱取对数后再进行傅里叶变换得到的结果
  - 系数：提取出的特征值，通常取12-13个系数
  - 是语音识别和声纹识别最常用的特征

语音压缩和MP3

- 70年代，德国教授迪特·塞策 （Dieter Seitzer）试图用ISDN电话线播放音乐，但速率只有128kbps，Brandenburg博士接受该任务。
- **Psychoacoustics（心理声学）与MP3**
  - **等响曲线**： 20-20kHz 范围里不同频率人耳听觉响度是不一样，如贝斯声音需要更大的音箱
  - **声音掩蔽**（Sound Masking）：两个频率的声音同时作用于人耳时，响度较高的频率的声音会影响响度较低的声音，使其不容易被察觉



等响曲线

语音与物联网

- **语音信号处理在物联网领域的应用**
  - **语音识别ASR（Automatic Speech Recognition）：说什么内容?**
  - **文语转换TTS（Text to speech）：文字转换成语音**
  - **身份识别：是谁在说话?**
  - **活体检测：区分发声体是人还是机器**
  - 人机对话：物联网移动、终端设备与人沟通
  - 语种识别：说的是什么语言?
  - 语音编码：高效的传输与存储?

Q：你能想到哪些语音类物联网应用?

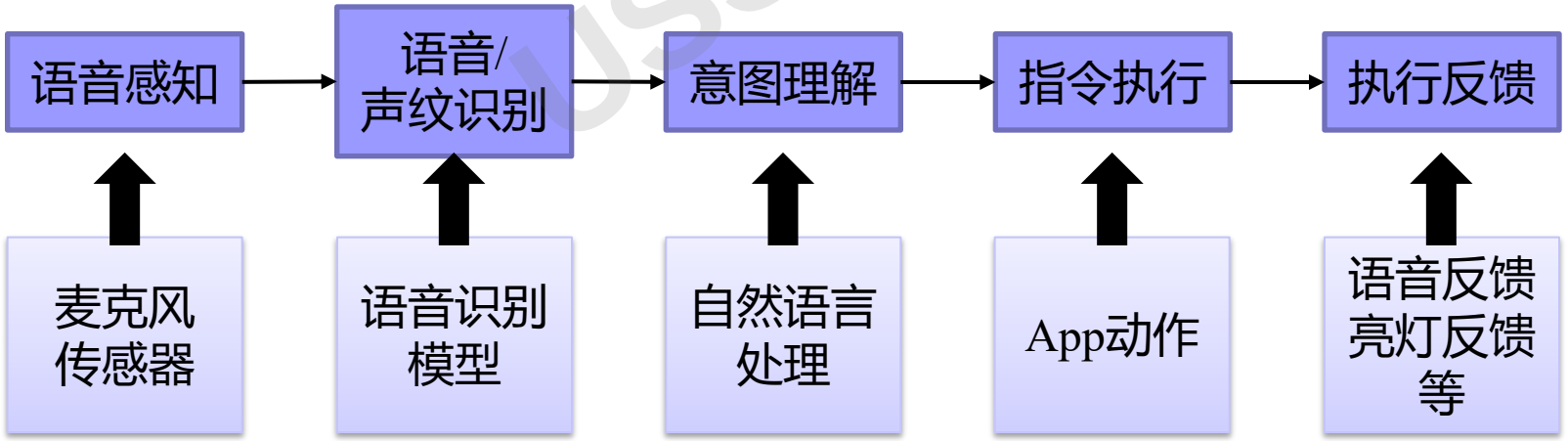
语音与物联网安全——智能语音系统

- **智能语音系统：**一种感知并利用语音信号的系统，它利用麦克风将自然界中人发出的声波转换成电信号，实现声音的收集，和**认证、控制、决策**等目的。
  - **人与终端交互：**
    - Google 语音识别
    - Apple Siri
  - **应用场景：**
    - 百科查询
    - 智能控制
  - **新兴应用**
    - 大模型
    - 具身智能

Q：为什么各大公司都在开发语音交互产品？

智能语音系统：工作流程

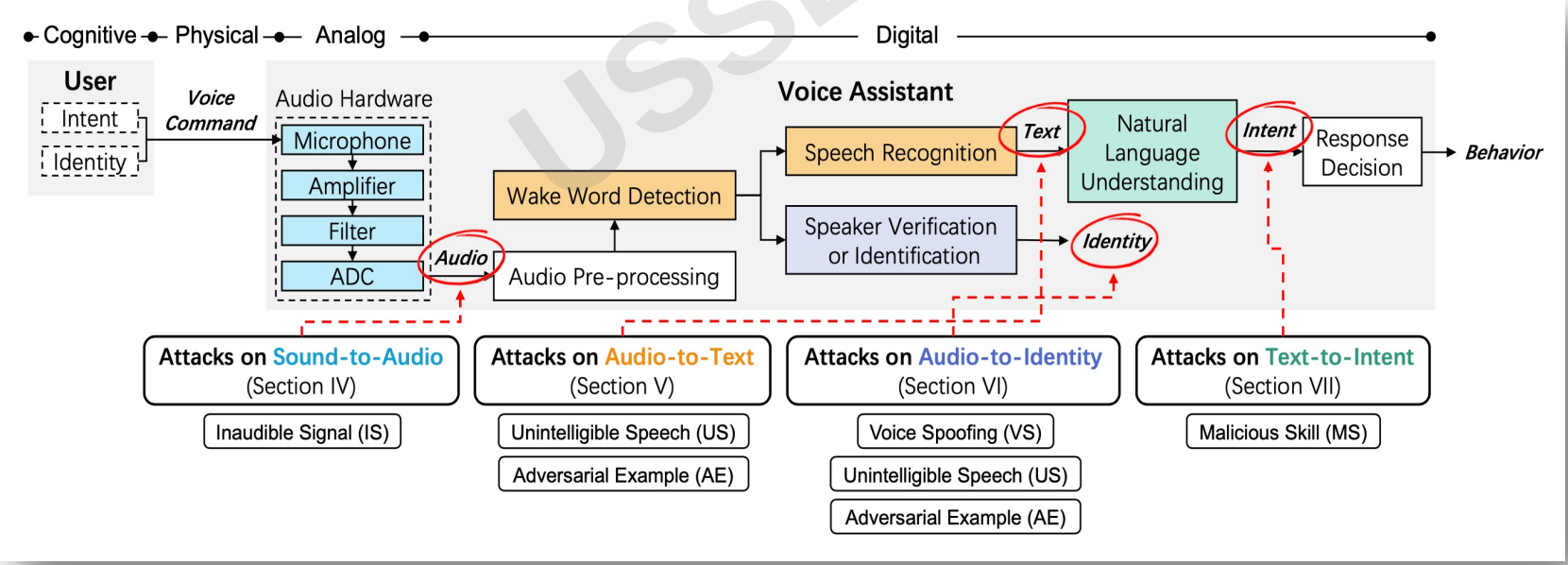
- **语音感知**：麦克风获取信号并进行信号处理，如放大、滤波、AD转换
- **语音/声纹识别**：将语音信号转换为文本数据/识别出说话人身份
- **意图理解**：将文本数据转换为可理解的指令内容
- **指令执行**：执行指令内容
- **执行反馈**：通过声音或者执行动作进行反馈



举例：通过Amazon Echo打开窗帘……

智能语音系统：工作流程及其脆弱性

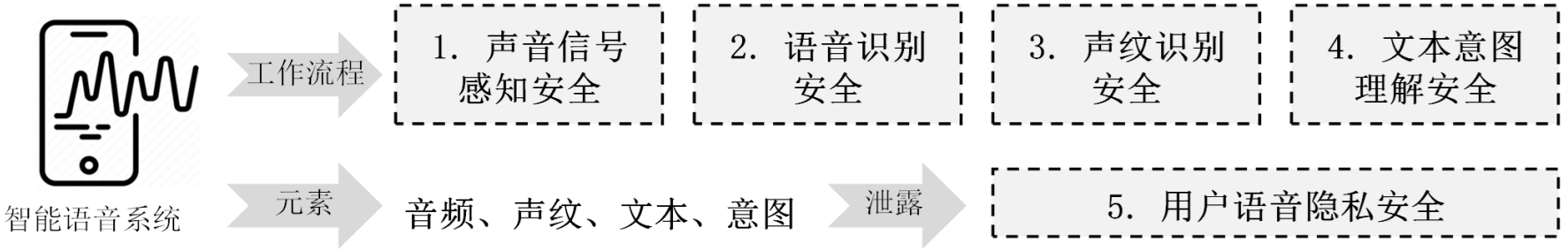
- **声波→语音信号→语音文本→语义理解/身份识别→意图执行**
  - 脆弱性1：**声音信号→语音信号**
  - 脆弱性2：**语音信号→文本转换**
  - 脆弱性3：**语音信号→身份识别**
  - 脆弱性4：**文本指令→意图理解**



A Survey on Voice Assistant Security: Attacks and Countermeasures, in ACM Computing Survey, 2022

物联网语音安全

- **定义：**语音在经过感知、传输、处理、识别、执行等从声音信号到数字信息的转换过程中，其语义和说话人信息具有安全性
- 根据语音在物联网中的应用场景，本课程将语音安全分为语音系统安全和用户语音隐私安全两类
- 其中语音系统安全包括：声音信号感知安全、语音识别安全、声纹识别安全、意图理解安全

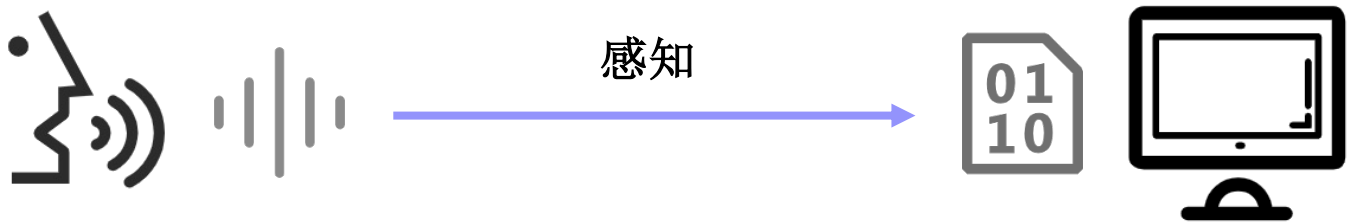


1. 语音感知安全

1. 声音信号感知安全

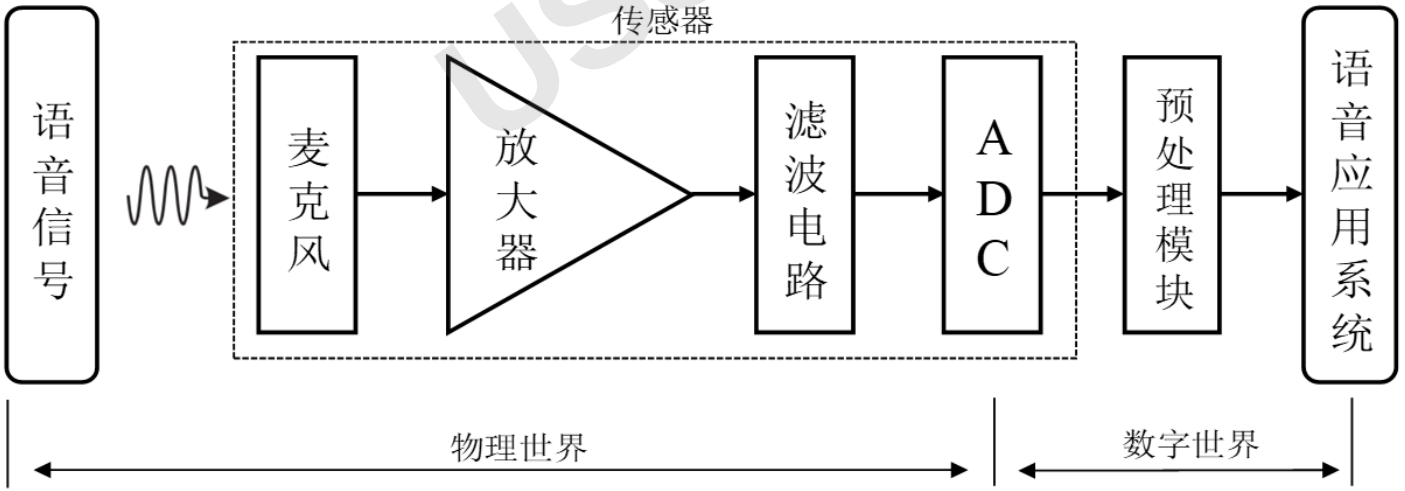
- **声音信号感知**
  - 定义：将**物理世界**的音频信号转换为**数字世界**的音频信息的过程
  - 声音信号感知过程一般需要麦克风传感器实现

- **声音信号感知安全**
  - 攻击者通过向传感器注入声音信号或其他物理信号，并利用传感器脆弱性（回顾传感器安全内容）影响后续声音信号处理，如语音识别和声纹识别



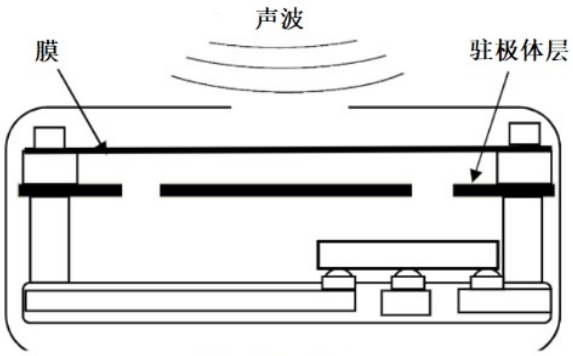
1. 声音信号感知安全

- **声音信号感知（背景知识回顾）**
  - 麦克风**传感器**：由麦克风换能器、放大器、滤波器、ADC等组成
  - 麦克风换能器将机械声波信号转换成模拟电信号，进而经过放大器放大、滤波电路进行滤波，最后经过模数转换器转换成数字电信号

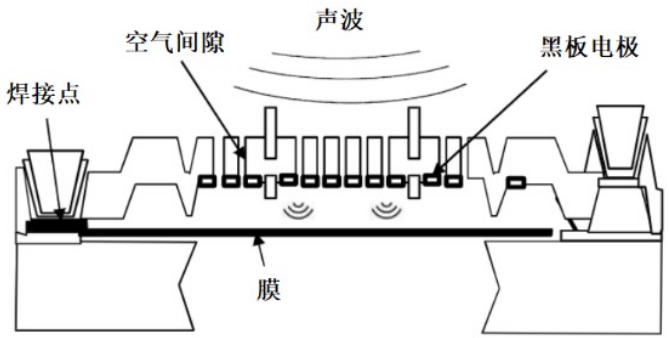


1. 声音信号感知安全

- **麦克风传感器原理**
  - 麦克风换能器一般具有一片可以感受声波的**薄膜**，该薄膜随声波变化发生相应的形变
  - 声音引起薄膜形变时，将改变薄膜与金属极板间的**距离**，从而影响电容值并产生**随声音变化的电压**



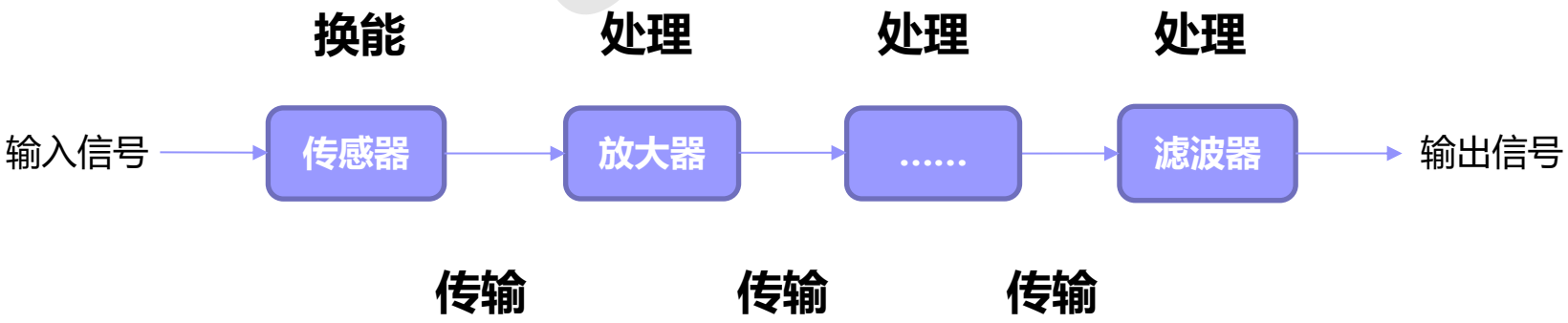
驻极体麦克风



MEMS麦克风

1. 声音信号感知安全

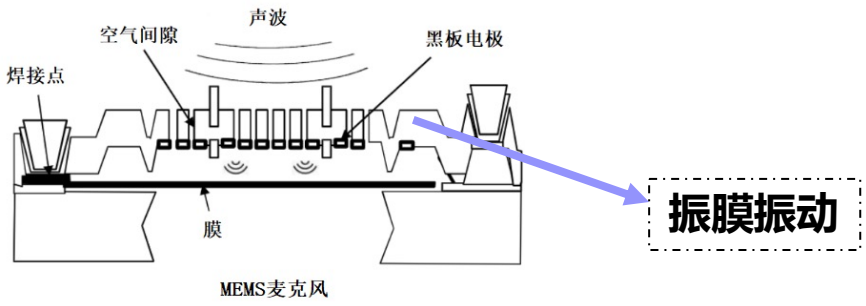
- **声音信号感知过程脆弱性**
  - 声音信号的感知过程主要存在以下两个方面的脆弱性
    - 信号**换能**过程中的脆弱性
    - 信号**处理**过程中的脆弱性
    - 信号**传输**过程中的脆弱性



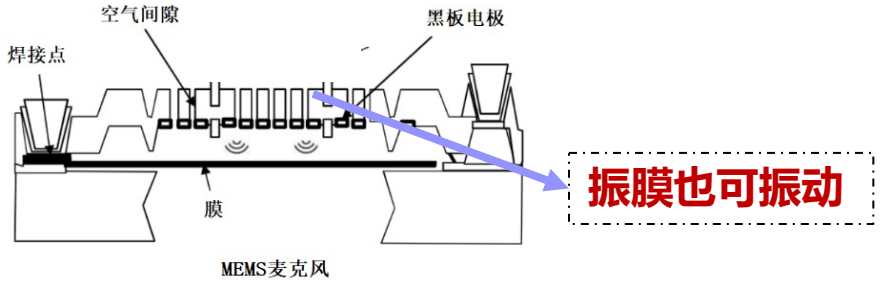
1. 声音信号感知安全

- **信号换能和处理过程中的脆弱性**
  - 由于硬件设计原因，麦克风换能器对机械振动类**物理激励敏感**，包括一些非功能设计输入的物理信号激励
  - 例如，麦克风传感器的换能器件对**超声波、激光信号**都具备敏感性，导致麦克风可以接收到此类攻击信号

声波



超声波、激光



1. 声音信号感知安全

- **信号传输过程中的脆弱性**
  - 电磁干扰：任何能够中断、阻碍、降低或限制通信电子设备有效性能的电磁能量
  - **导线天线效应**：通过电磁干扰来改变传感器工作链路中传输的信号，从而改变传感器模块输出的测量值
  - 案例：GhosTalk（自行回顾上一章）

1. 声音信号感知安全：LightCommand

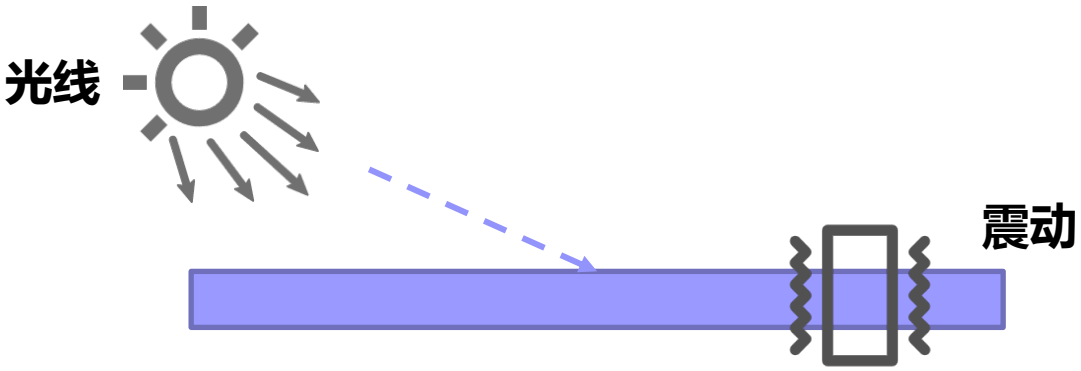
- **LightCommand**
  - 基于**光声效应**，面向MEMS麦克风传感器的注入攻击
  - 攻击者通过使用AM调制之后的激光信号注入MEMS麦克风模块孔径，向目标麦克风注入**任意音频信号**

Sugawara, Takeshi, et al. “Light commands: Laser-Based audio injection attacks on Voice-Controllable systems , USENIX Security’20

https://lightcommands.com/

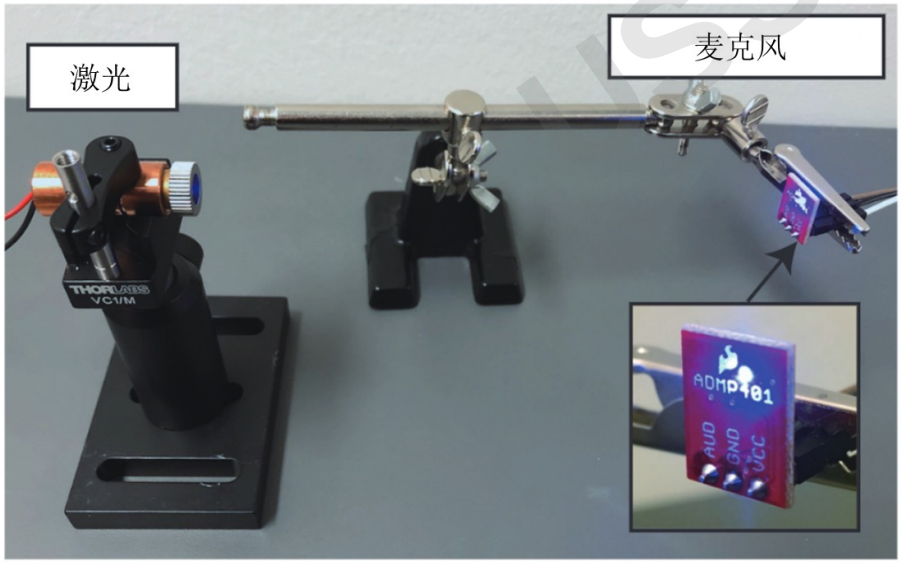
1. 声音信号感知安全：LightCommand

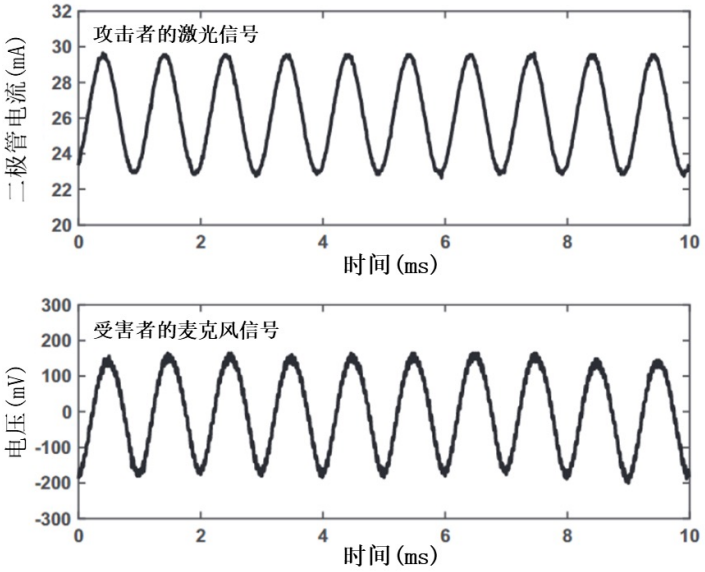
- **光声效应**
  - 也称为光热声效应。指物体在周期性变化的**光照下产生声音信号**的现象。
  - 物体在光的照射下能量增加，局部聚集的能量以热的形式释放并引起**周围物质的震动**，震动频率由光信号的频率决定，震动的强度由物体的材料、几何形状等性质有关。



1. 声音信号感知安全：LightCommand

- **Light Commands过程**
  - 攻击者将语音指令转换成驱动电流信号，电流变化**改变激光信号强度**，从而实现麦克风注入恶意语音指令



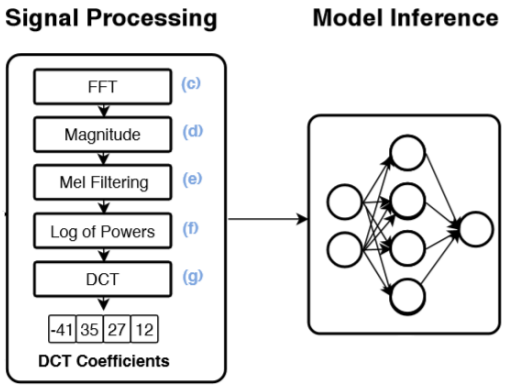


2. 语音识别安全

2. 语音识别安全

- **语音识别过程包括**
  - **特征提取**：提取语音信号时频域特征，如MFCC、Embedding等
  - **模型推理**：将特征输入到语音识别模型中，输出识别结果

- **脆弱性分析**
  - **特征**：提取的特征不准确或者出错，导致识别错误
  - **模型**：语音识别模型被攻击，如各类对抗样本、数据中毒攻击等
  - 结果：**耳听不为实**



2. 语音识别安全——攻击分类

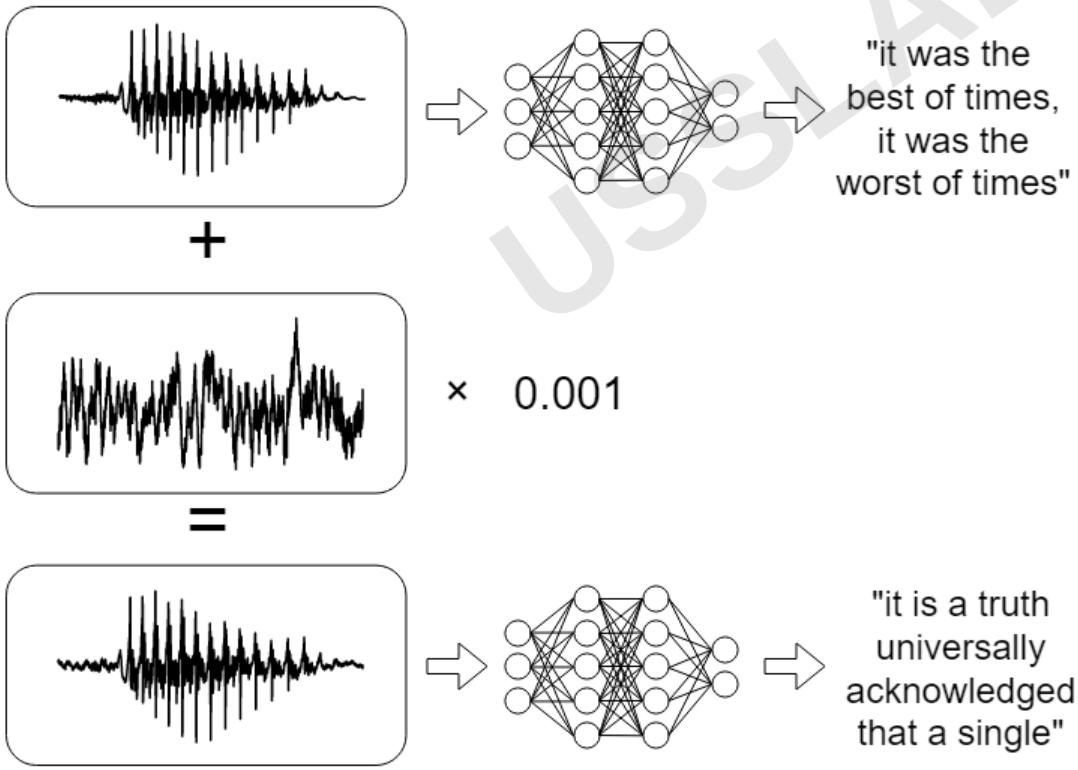
- **1. Normal speech攻击**
  - 人可以理解攻击语音（如攻击者直接说出恶意指令）
- **2. Unintelligible攻击**，通常对抗样本形式出现：
  - 人不能理解攻击语音，但是机器可以理解，大部分是基于算法脆弱性实现的语音对抗样本，如hidden voice command [1]
- **3. Imperceptible攻击**，通常对抗样本形式出现
  - 人和机器对攻击语音的理解不同（如一首歌中隐藏攻击指令）
- **4. Inaudible攻击（如海豚音攻击）**
  - 攻击指令完全不可听，不被人察觉。如海豚音攻击等

上述四类攻击可听性/可察觉性从上到下依次减小。

[1] Nicholas Carlini, et. al, hidden voice commands, USENIX Seucirty'16

2. 语音识别安全——对抗样本攻击

- **语音对抗样本**：攻击者对语音做特定处理（添加特定噪声），使语音识别系统识别结果和人识别结果产生不一致的现象（具体内容AI安全专题会进一步介绍）



语音对抗样本攻击——威胁模型

- **攻击目标：**
  - **无目标**：使识别模型的预测结果与正常音频的识别结果不同，例如让正常音频的“lock”识别错误
  - **有目标**：预测结果可以由攻击者指定，例如让“lock”识别成“open”

- **攻击者要求：**
  - **白盒**：攻击者完全掌握待攻击目标模型的知识
  - **灰盒**：攻击者只掌握待攻击目标模型所有组件的一个子集
  - **黑盒**：攻击者只知道模型的输入输出信息以及目标模型的作用

方法：Carlini & Wagner Attack

- **C&W攻击**：通过在音频上添加难以觉察的扰动，使语音识别模型识别结果出错

$$\text{minimize } dB_x(\delta)$$
$$s.t.\, C(x+\delta) = t,$$
$$x + \delta \in [-M, M]$$

$\Longrightarrow$

$$\text{minimize } |\delta|_2^2 + c \cdot \ell(x+\delta, t)$$
$$\text{s.t. } dB_x(\delta) \le \tau$$

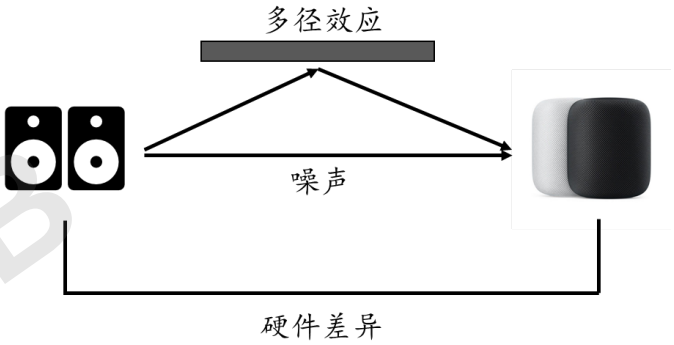
将不可导的约束转换为可导的损失函数$\ell(\cdot)$

- **攻击对象: 百度Deepspeech 0.4**
  - **原句**：无识别结果
  - **对抗样本**：speech can be embedded in music
  - **原句**：It is time for class
  - **对抗样本**：open the door

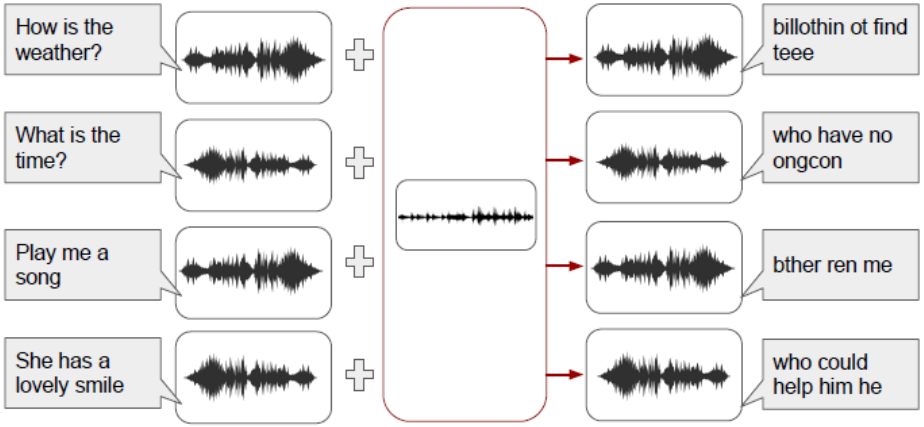
Nicholas Carlini, David Wagner Towards Evaluating the Robustness of Neural Networks, IEEE S&P'17

语音对抗样本攻击——现实挑战

- **通过空气传播（Over the air, OTA）**
  - 多径效应
  - 信号衰减
  - 环境噪声



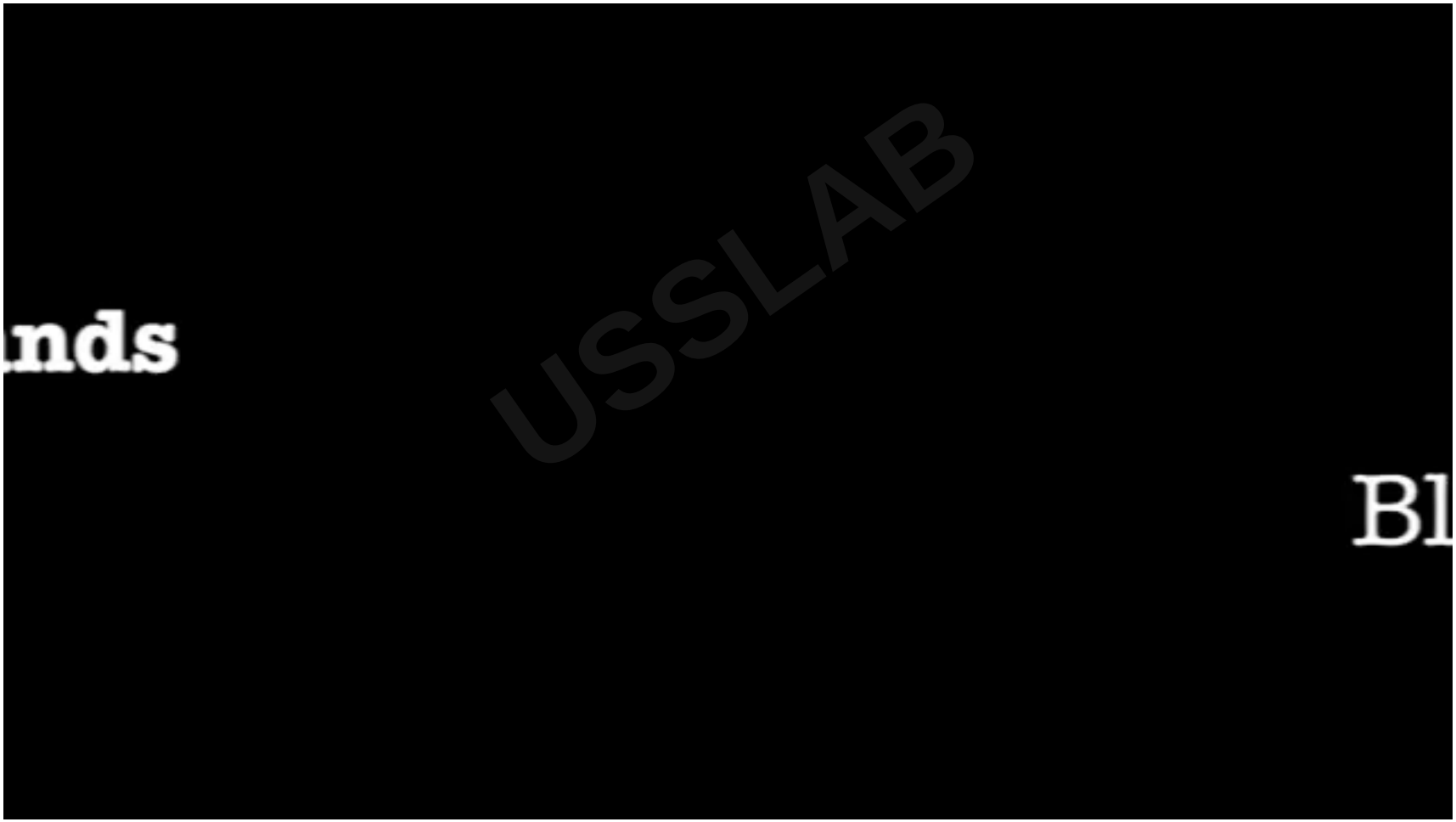
- **通用对抗样本**
  - 不依赖输入
  - 不依赖同步



你认为通用对抗样本可行吗?

2. 语音识别安全： Unintelligible攻击

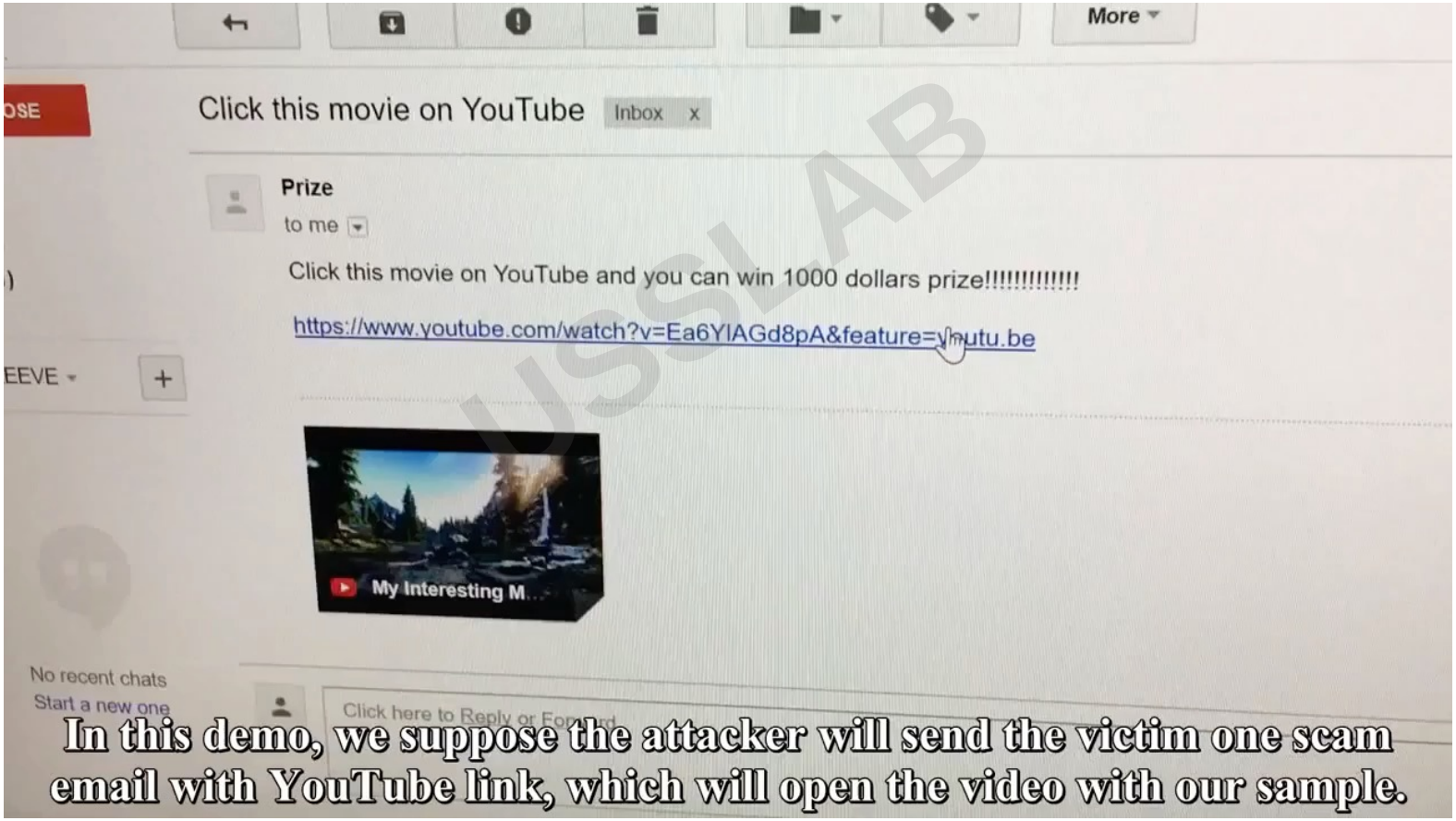
案例 - Hidden Voice Commands



Ref: http://www.hiddenvoicecommands.com/home

2. 语音识别安全：Imperceptible攻击

案例1 - CommanderSong



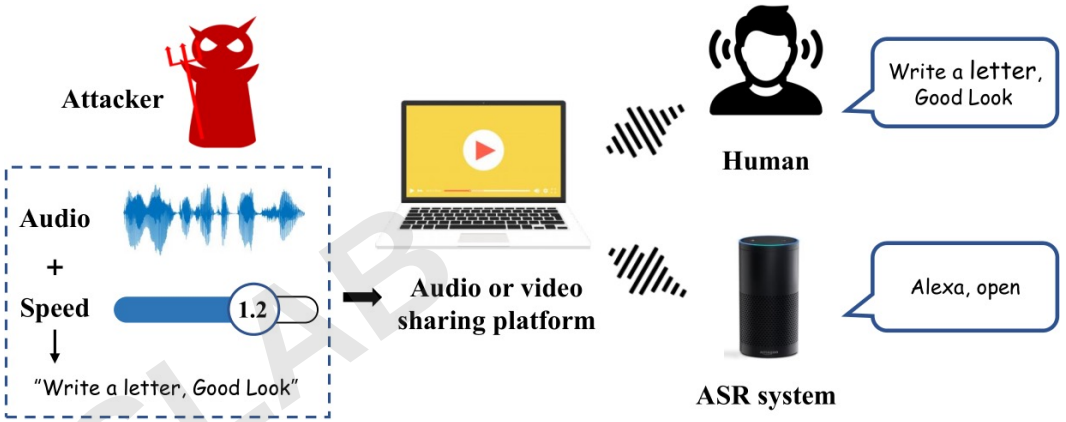
Yuan, Xuejing, et al. CommanderSong: A systematic approach for practical adversarial voice recognition. USENIX security 18

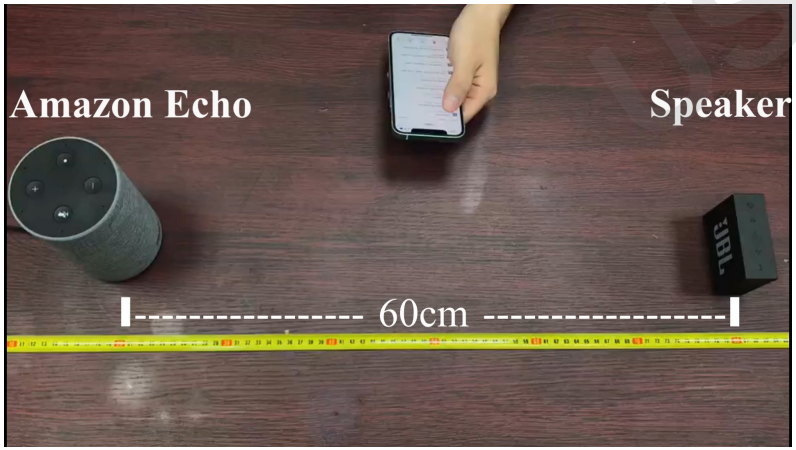
Ref: https://www.youtube.com/watch?v=t5-vMJDFr8E

2. 语音识别安全：Imperceptible攻击

案例2 – TSMAE倍速攻击

TSMAE攻击 [1]：通过改变部分音频帧的速率，使语音识别结果错误





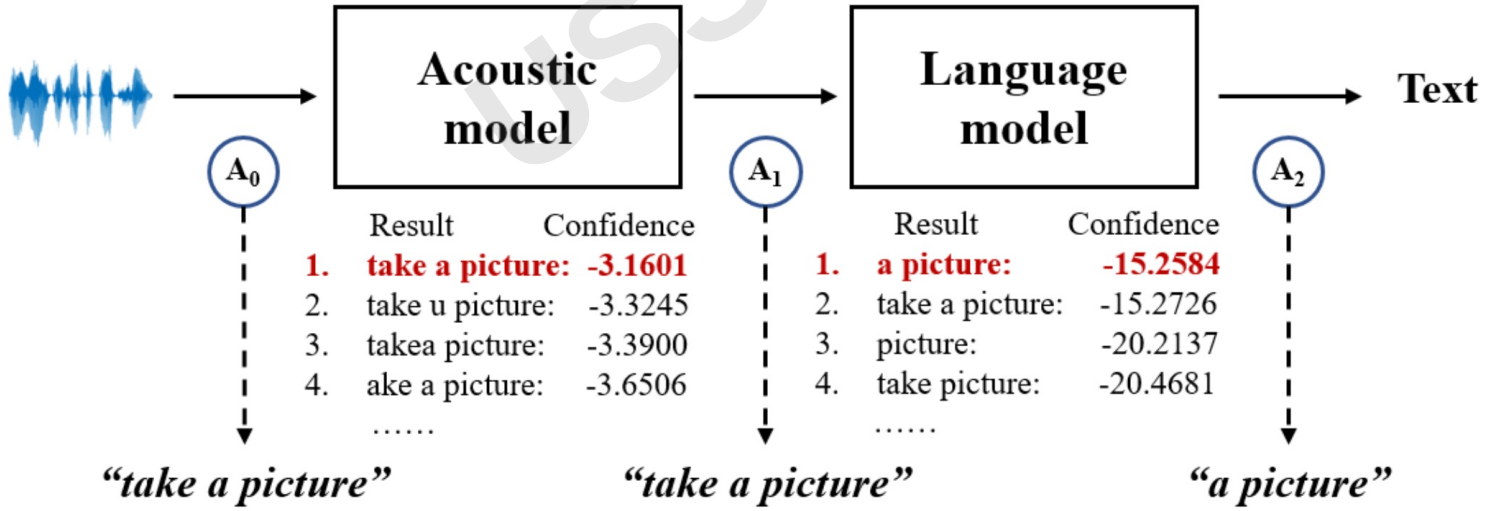
- **直接注入攻击: 百度Deepspeech**
  - **原句**：Eurpoean the duke of York
  - **变速后误识别为**：**open the door**

- **经空气传播攻击: 亚马逊Echo**
  - **原句**：Biden is an elector
  - **变速后误识别为**：**Alexa**

[1] Xiaoyu Ji, et. al, "Watch Your Speed: Injecting Malicious Voice Commands via Time-Scale Modification", IEEE ***TIFS***, 2023.

TSMAE攻击原理分析

- **语音识别模型：**
  - Acoustic model：语音到字母
  - Language model：字母到符合语义、语法的单词
- **结果**：速度变化导致字母/单词增加、丢失、变换



https://sites.google.com/view/tsmae/homepage

3. 声纹识别安全

声纹识别概述

- **什么是声纹**
  - 声纹是生物识别特征的一种，是从声音信号提取的可以作为说话人的表征和标识的语音信号特征
  - 声纹有两个特性：
    - **稳定性**：成年以后，人的声音可保持长期相对稳定不变
    - **特定性**：无论讲话者故意模仿他人声音和语气，还是耳语轻声讲话，即使模仿得惟妙惟肖，其声纹却始终不变

- **声纹识别**
  - 声纹识别（Voiceprint Recognition）是根据说话人的声波特性进行身份辨识的技术
  - 也称为说话人识别（Speaker Recognition）

声纹识别的优势

- 蕴含声纹特征的语音获取方便、自然，声纹提取在不知不觉中完成，因此使用者的接受程度也高；
- 语音采集装置造价低廉，只需手机或麦克风即可，无需特殊的设备
- 与指纹、人脸相比，声纹更适合于远程身份认证

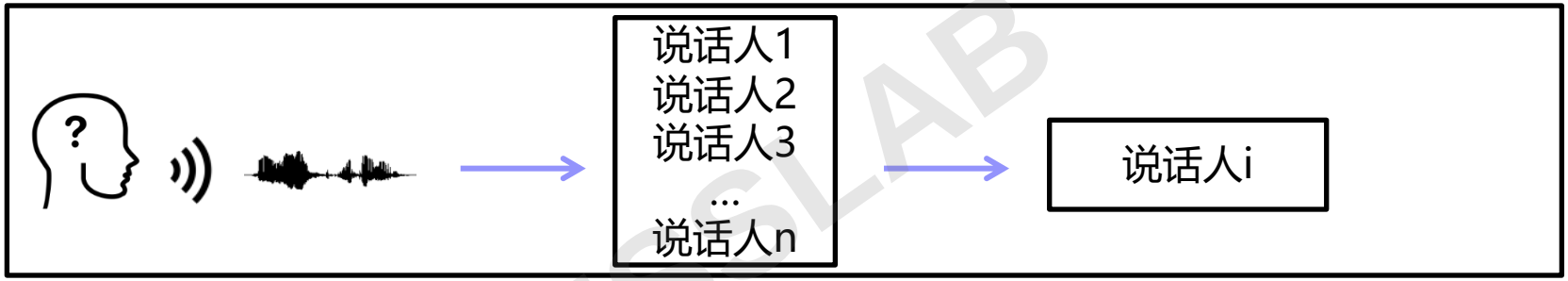
智能设备认证

金融身份认证

国防安全、公安技侦

声纹识别分类——根据识别任务

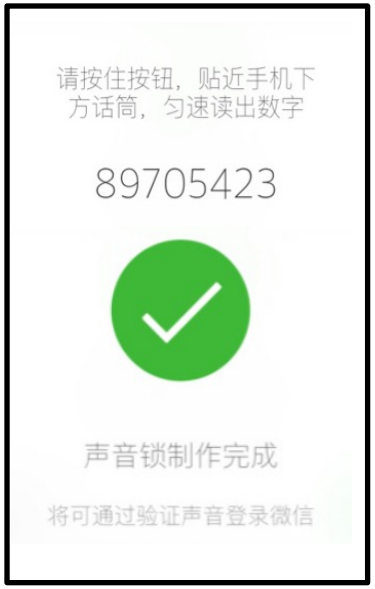
- **说话人辨认(Speaker Identification)**：判断某段语音是若干人中的哪一个所说的，是“**多选一**”问题



- **说话人确认(Speaker Verification)**：确认某段语音是否是指定的某个人所说的，是“**一对一判别**”问题

声纹识别分类——根据文本内容

- **文本相关**：要求用户按照规定的内容发音，每个人的声纹模型逐个被精确地建立；识别也必须按规定的内容发音
  - 系统需要用户配合，若用户的发音与规定的内容不符合，则无法正确识别该用户



- **文本无关**：不规定说话人的发音内容，模型建立相对困难，但用户使用方便，可应用范围较宽

声纹安全案例

- **新闻报道**

根据华尔街日报，犯罪分子通过商业化的人工智能语音生成软件，成功模仿并冒充一家英国能源公司的德国母公司CEO，来欺骗其多位同事和合作伙伴，一天内多次诈骗并转移资金，使得该公司损失220,000欧元（约合173万元人民币）。

声音可以被伪造


THE WALL STREET JOURNAL.

English Edition | October 30, 2019 | Print Edition | Video

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

BREAKING NEWS Pace of U.S. economic growth slowed slightly to 1.9% in third quarter as business investment declined, though consumer spending kept growth on track

SHARE

AA TEXT

PRO CYBER NEWS

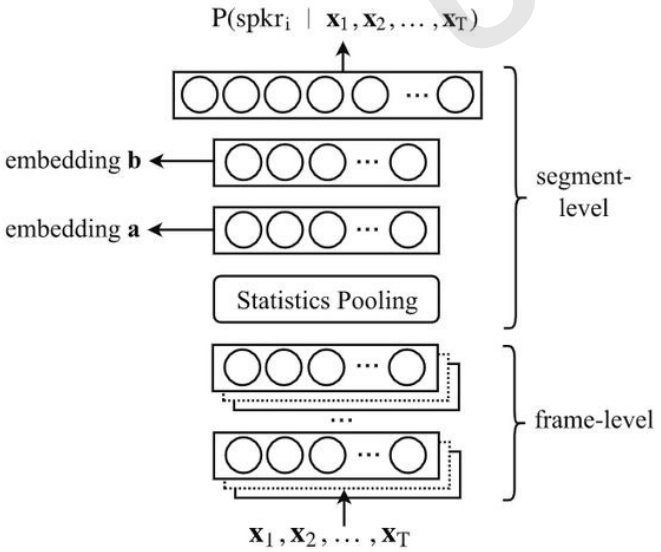
Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

新闻来源 ： https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402?sa=X&ved=2ahUKEwjuzKvwlYTmAhXMJzQIHbZkDPwQ9QF6BAgKEAI

基础知识：声纹识别模型DNN

- 随着深度学习的发展，声纹识别技术如基于DNN的X-Vector算法将声纹特征表示为Embedding，并用分类层来实现识别
  - **特征提取层：**
    - **帧级特征：**通过提取DNN层提取语音的帧级（约20-30ms）特征
    - **段级特征：**通过统计池化将音频帧级特征的均值和标准差连接作为段级特征
  - **分类层：**通过前馈网络将段级特征分类到说话人身份



基于DNN的声纹识别技术，通过深度神经网络提取speaker embedding（特征向量），并基于概率统计学方式进行建模

声纹识别安全概述

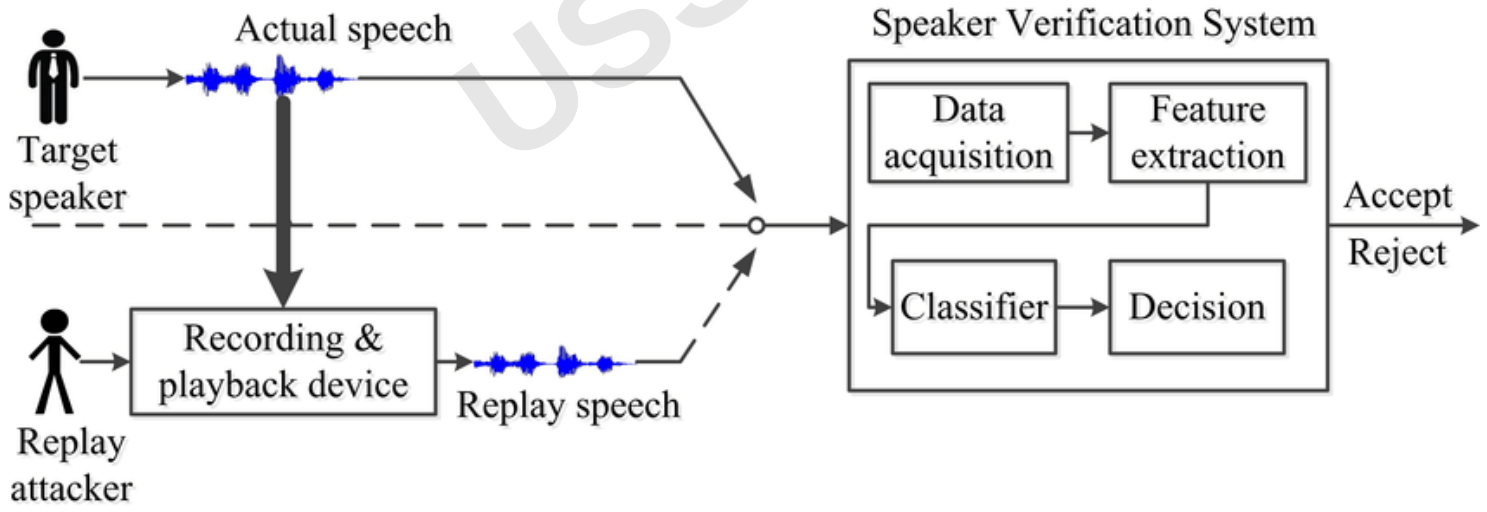
- **目标：**欺骗声纹识别系统，令其认为攻击者是合法的注册用户
- **分类：**包括扬声器重放被模仿者的声音、语音对抗样本等

- 具体技术：
  - **语音重放攻击**
  - **声纹合成攻击**
  - **声纹转换攻击**
  - **人为模仿攻击**

声纹识别安全——重放攻击

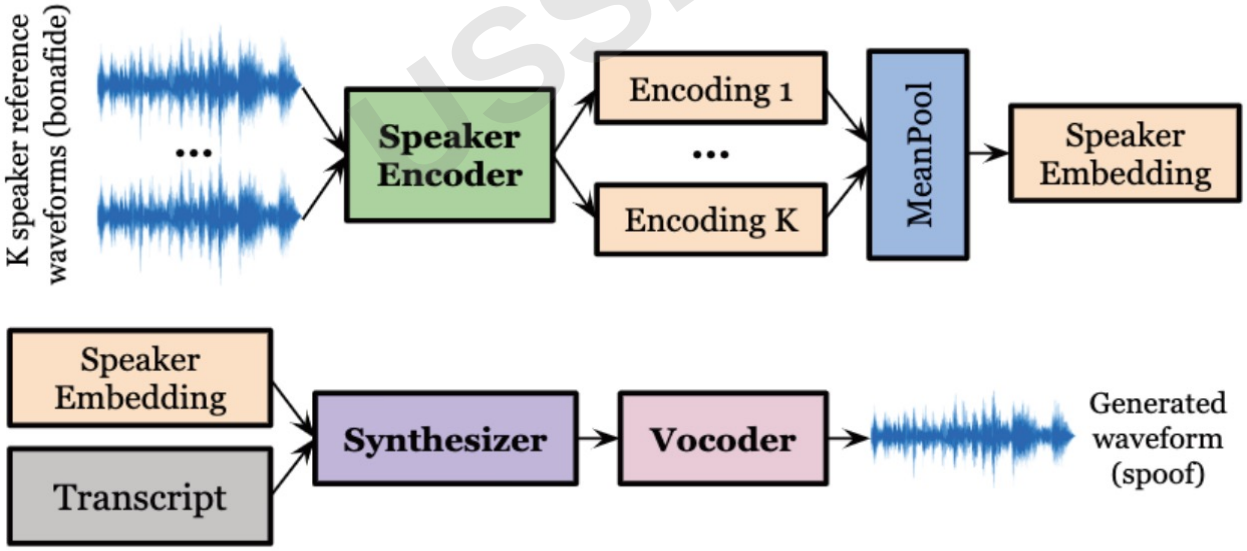
- 重放攻击：**重放已录制的受害者语音样本以欺骗声纹识别系统**，包含连续的语音记录或从多个语音记录中提取并串联而成

- 重放攻击技术上易于执行且无需复杂的语音处理技术，却可以有效欺骗现有的声纹识别系统。



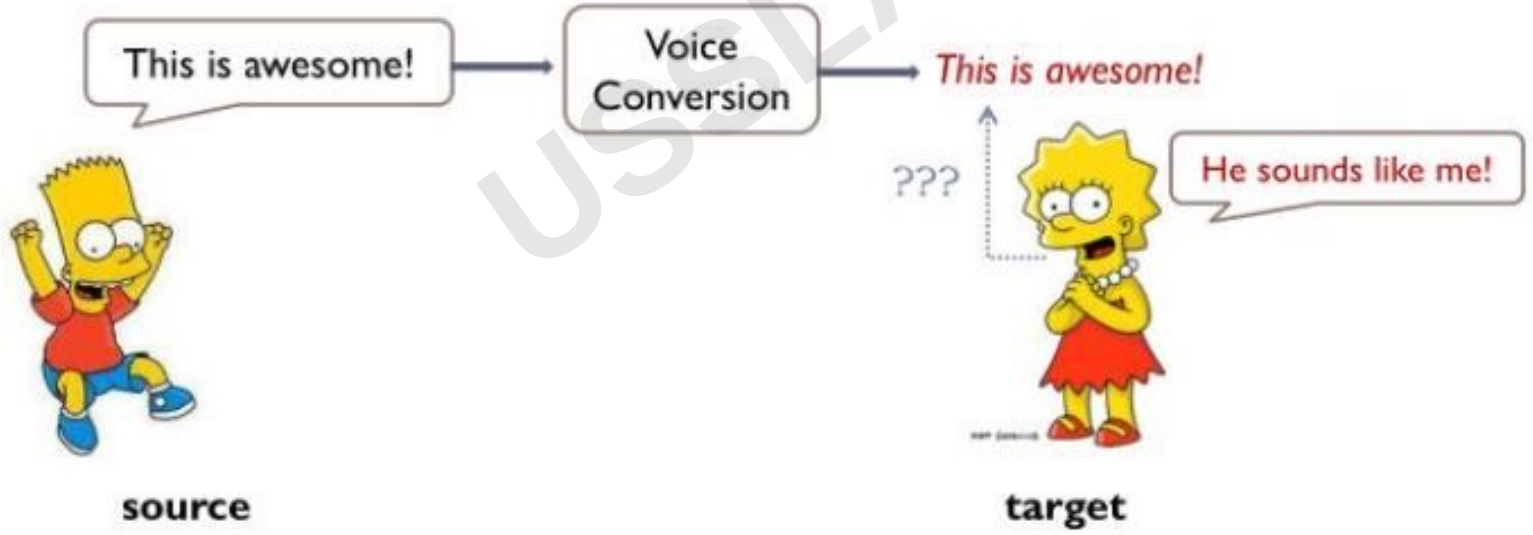
声纹识别安全——声纹合成攻击

- 声纹合成：将**任意文本生成目标说话人语音声纹的技术**
- 与文本到语音（Text-to-Speech，TTS）的合成不同，声纹合成攻击需要生成特定说话人声音以骗过声纹识别
- 过程包括：首先基于目标说话人语料，构建该说话人的表征，再进行目标说话人的声码器建模。



声纹识别安全——声纹转换攻击

- 声纹转换：将特定说话人（如攻击者）的语音，通过技术处理转换成其他具有合法用户语音内容的语音



声纹识别安全——人为模仿攻击

- 人为模仿攻击是指攻击者在没有计算机技术辅助情况下模仿目标说话人的声音（音色、音调等），类似“口技”
- 尽管人为模仿与其他欺骗方法相比更容易执行，但其攻击效果主要取决于冒名顶替者的声纹模仿能力，难以进行大规模评估

声纹识别攻击——对抗样本攻击

- 和语音识别对抗样本攻击类似，但是声纹对抗样本有其独特问题与挑战：
  - **分数阈值**：大多数声纹识别系统对说话人的验证音频打分，并用阈值决定是否与注册人匹配，攻击成功需要让模型输出高于阈值的分数，然而黑盒系统中无法输出分数反馈给攻击
  - **性别判断**：由于男性和女性声音之间的显著差异，性别间的攻击比性别内的攻击更复杂

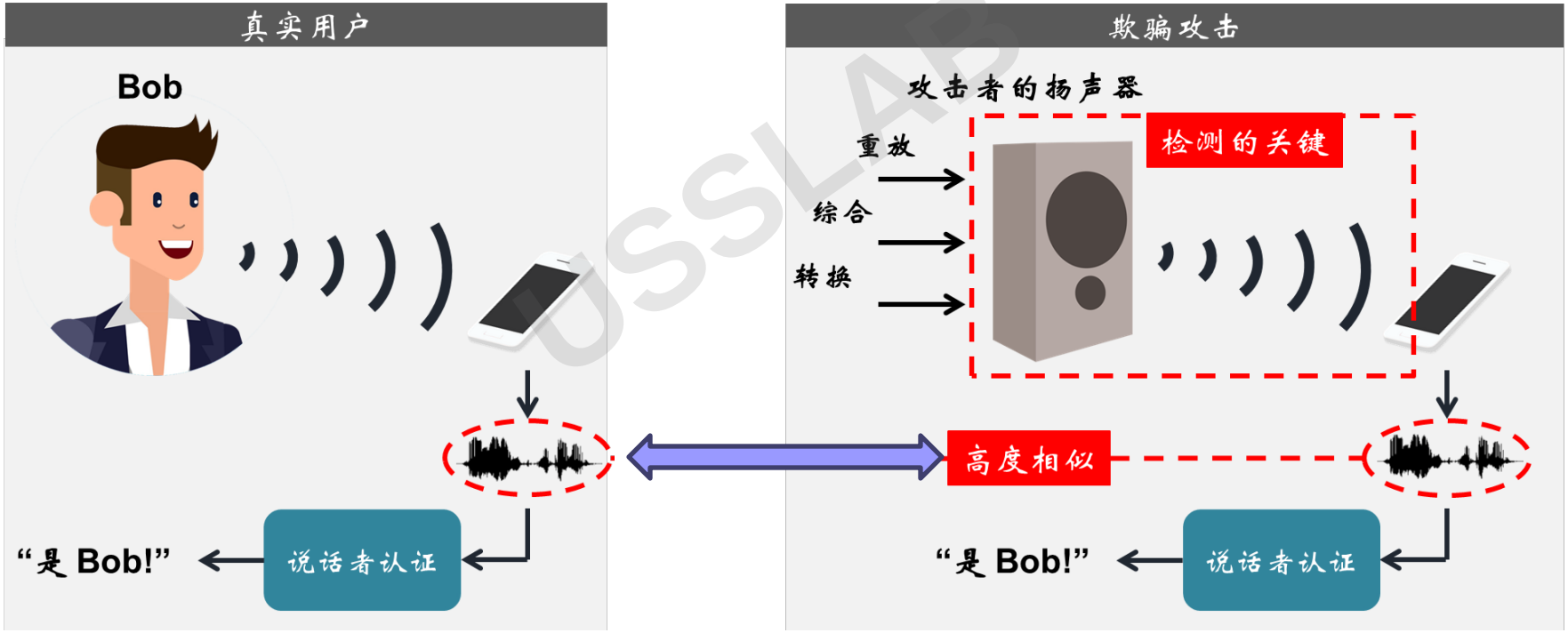
- 针对同时满足上述两类问题的黑盒声纹识别系统，实现对抗样本攻击是具有挑战性的

声纹识别安全——防御技术

- 重放攻击检测 - CaField
- 声纹安全度量 - ProleScore

防御案例 —— CaField

- 基于声场的重放攻击检测：
  - 场景：一个人和一个扬声器同时说话，如何区分是人还是扬声器？

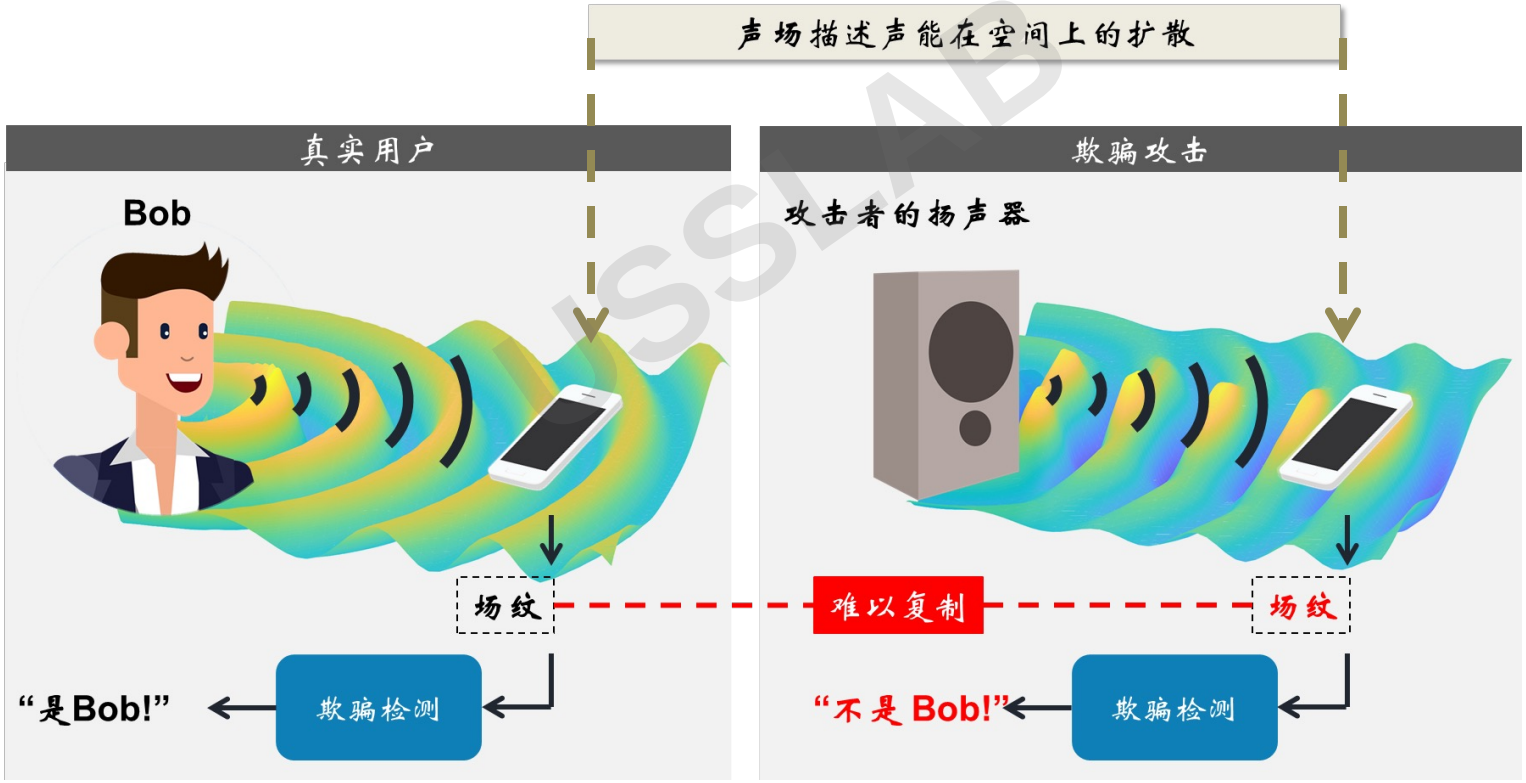


普通的声纹识别无法判断“说话人”是人还是音箱

Chen Yan, et. al, “Catcher in the Field: A Fieldprint based Spoofing Detection for Text-Independent Speaker Verification”, ACM CCS’19

防御案例 —— CaField

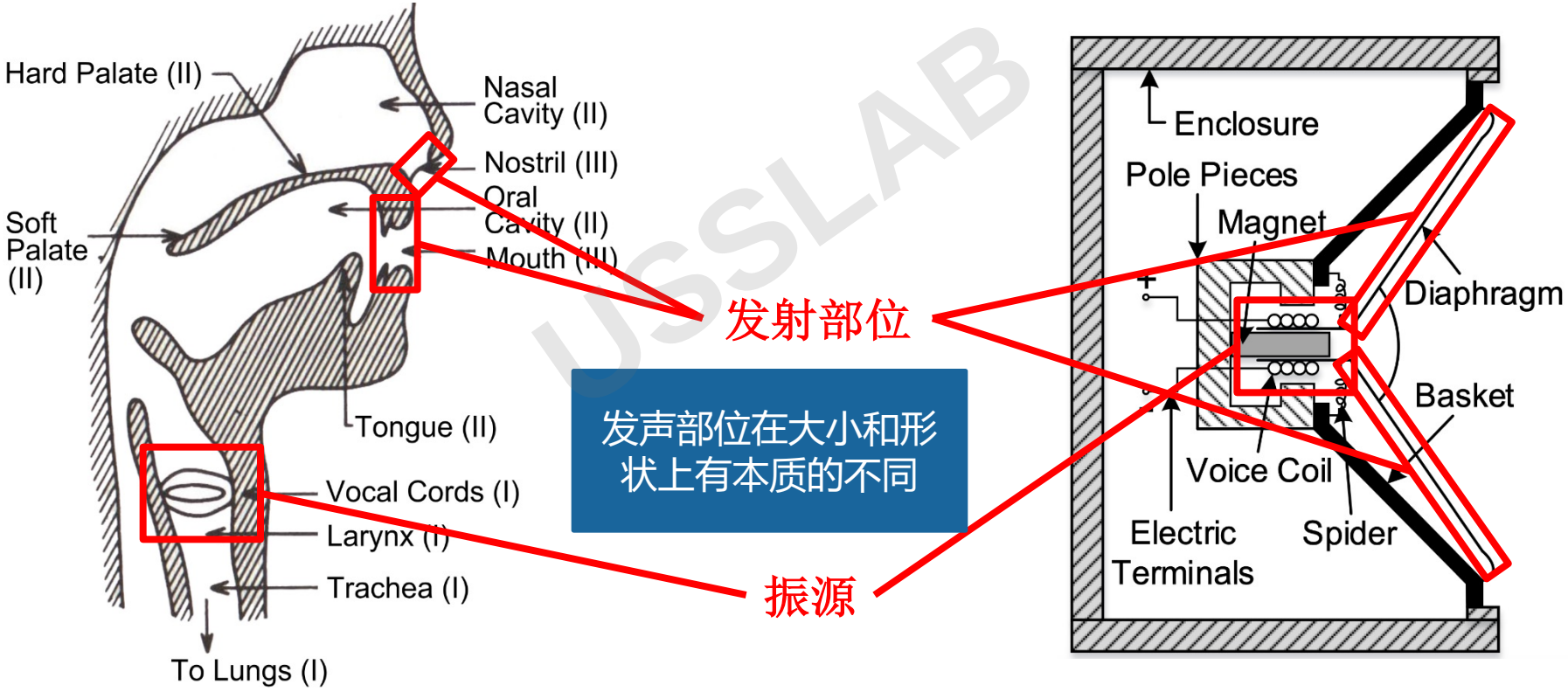
- 声场的优点：
  - 场景：一个人和一个扬声器同时说话，如何区分是人还是扬声器？



利用声场则可以判断“说话人”是不是人，即可以进行活体检测

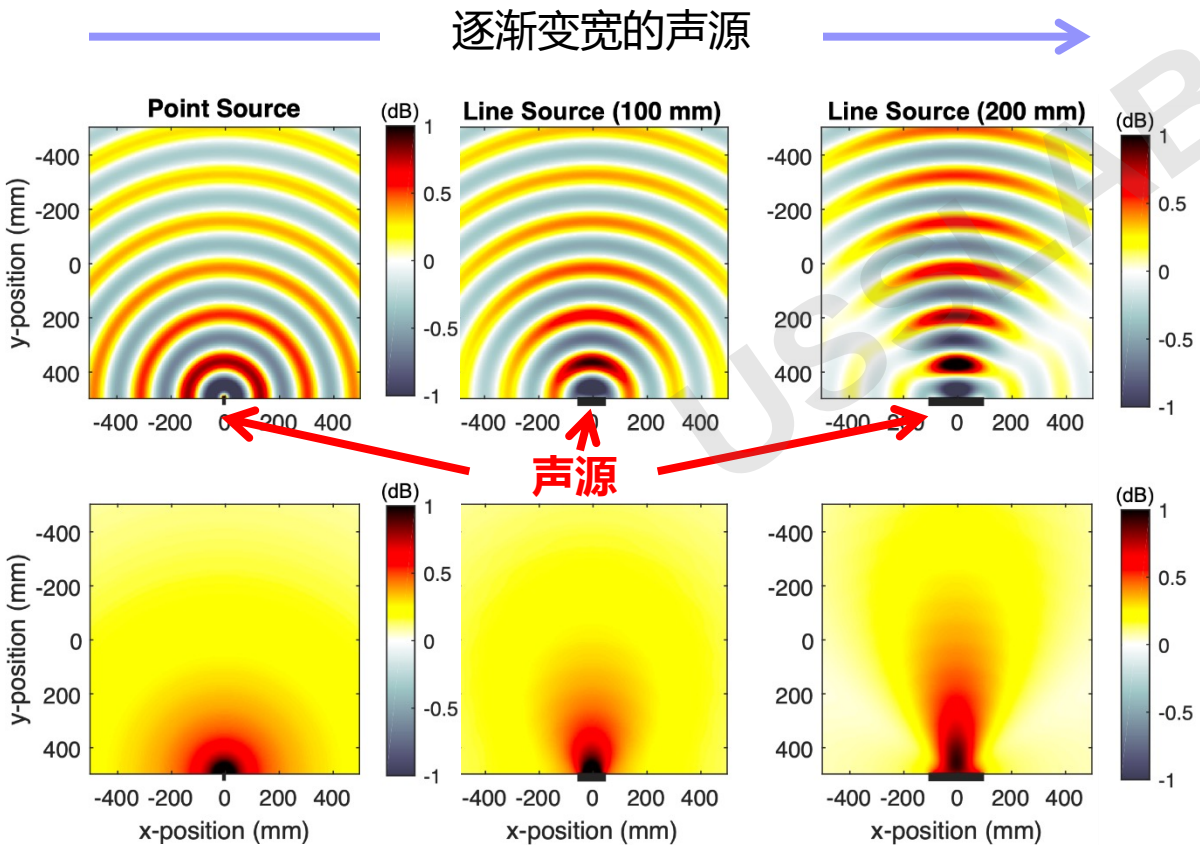
防御案例 —— CaField

- 运用声场区分真实用户和欺骗攻击者



防御案例 —— CaField

- 运用声场区分真实用户和欺骗攻击者



更宽的尺寸 → 更有方向性

作为声源，人和扬声器在尺寸、构造上有什么不同？

声源形状对声场的影响

防御案例 —— CaField

- 只用手机（而不需要额外设备）就能从声场中提取声纹
- 智能手机上麦克风数量一般至少2个
- 测量2个位置的声场能量差异

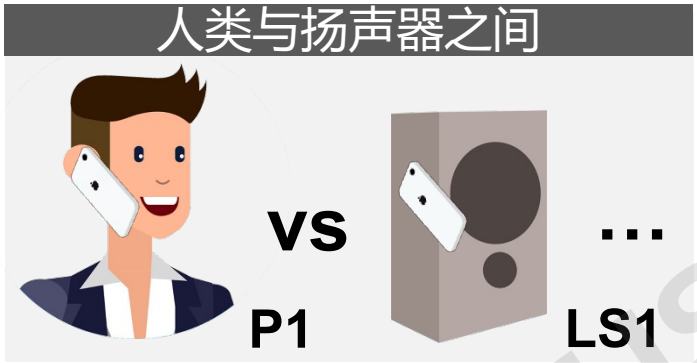
$$S_R(p_1, p_2, f) = \frac{s(p_1, f)}{s(p_2, f)}$$

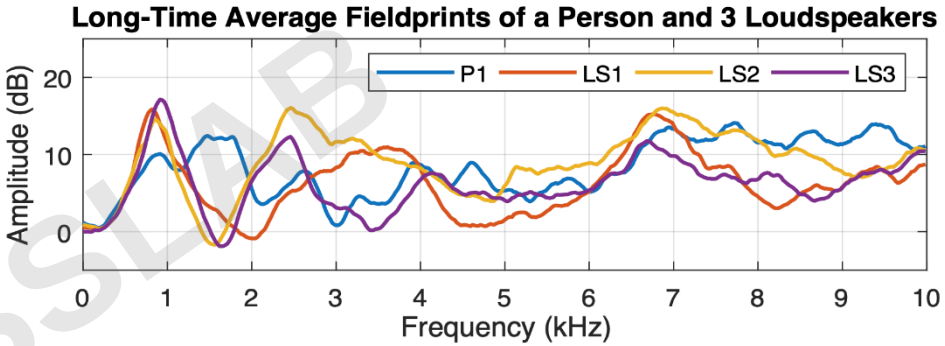
$s(p_1, f)$ ← 麦克风1处的声场强度

$s(p_2, f)$ ← 麦克风2处的声场强度

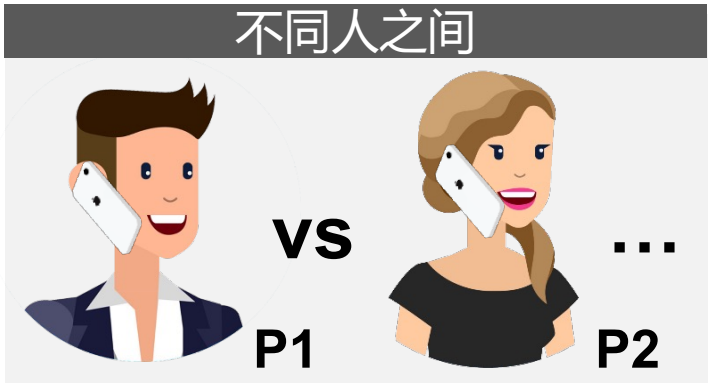
防御案例 —— CaField

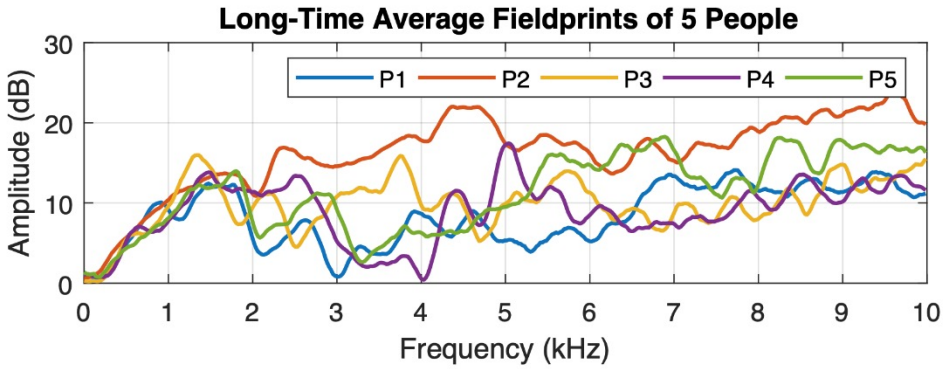
- 区别性（distinctiveness）





一个人和三个扬声器的长时间平均场纹对比

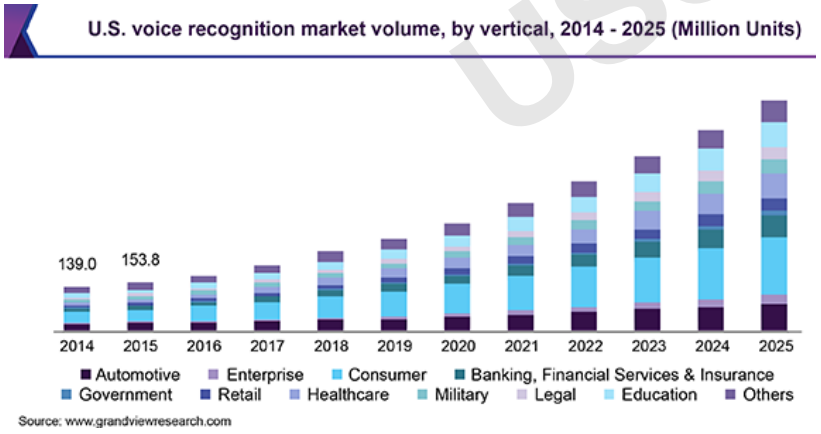




五个人的长时间平均场纹对比

声纹安全理论分析：PROLE Score

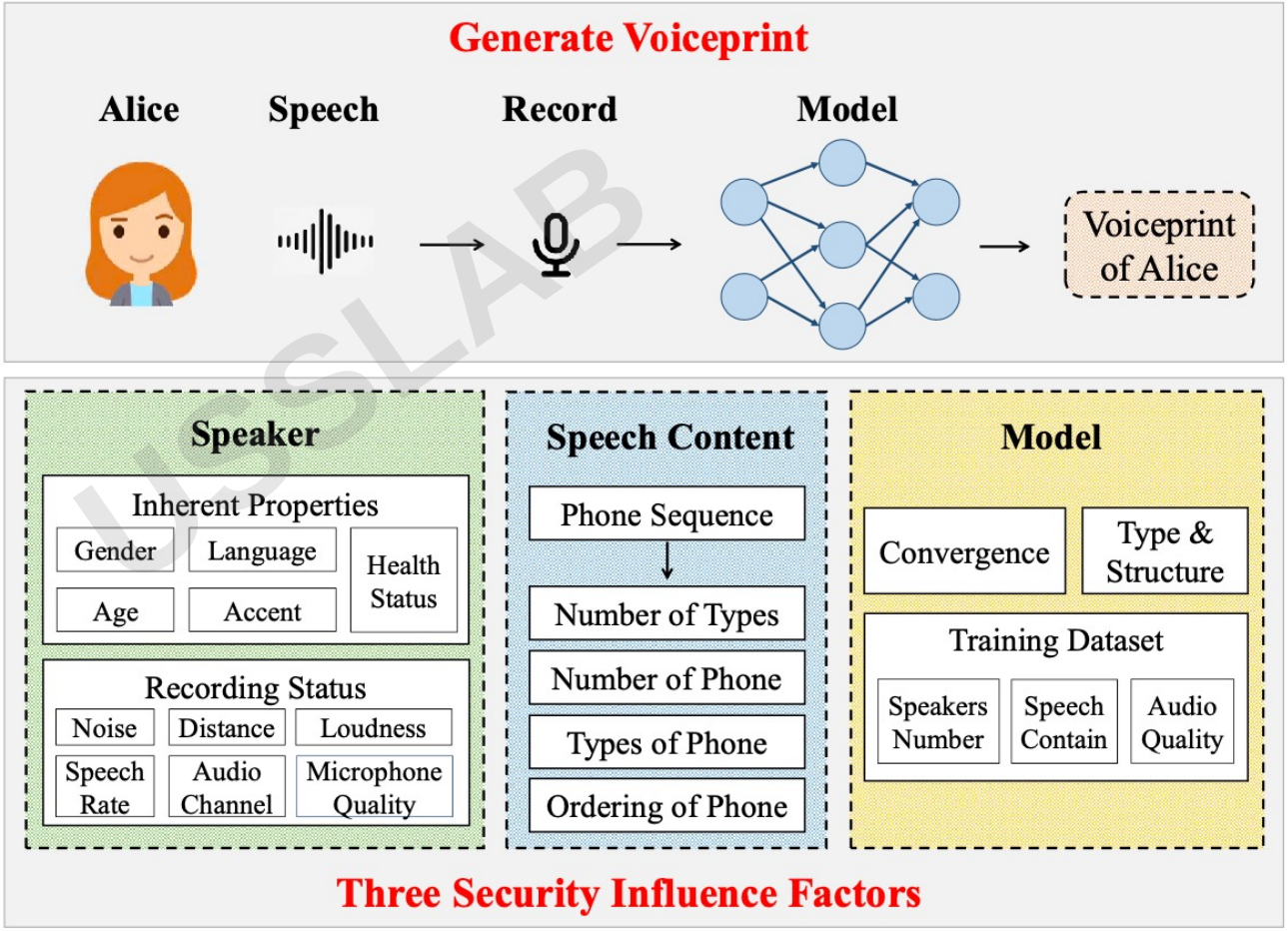
- 指纹识别算法的冲突概率大概是$1/10^5$，那么声纹呢？
  - 对于唤醒词来讲，“小欧小欧”、“小度小度”、“Hey, Siri”、“OK, Google”……哪个更安全？
  - 为什么没有人用“啊啊啊啊啊”作为唤醒词，或者为什么没有人用“八百标兵奔北坡”☹
- **核心问题**：现在的声纹认证系统有多安全，**声纹安全和哪些因素相关**



Ruiwen He, Xiaoyu Ji*, Xinfeng Li, Yushi Cheng, Wenyuan Xu, "OK, Siri" or "Hey, Google": Evaluating Voiceprint Distinctiveness via Content-based PROLE Score *USENIX Security,* 2022.

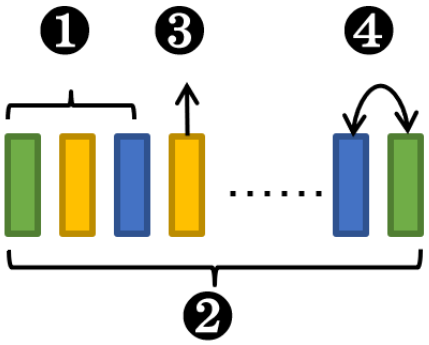
声纹安全理论分析：PROLE Score

- **用户**
  - 身份
  - 年龄
  - 性别
  - 音色
  - ……
- **内容**
  - 长短
  - 元音、辅音
  - ……
- **模型**
  - i-vector
  - DNN
  - ……



声纹安全理论分析：PROLE Score

- **声纹的区分性度量指标：PROLE Score**
  - 关注说话语音内容
  - 语音内容可以从音素角度的拆解为4个方面：
    音素丰富度、音素长度、音素种类、音素顺序



❶ Richness R → **PROLE Score**
(the number of phone types)

❷ Length L
(the number of phones)

❸ Element E
(the specific type of phones)

❹ Order O
(the sequential relationship among phones)

$$S = f(C) = f(R, L, E, O) \quad (1)$$

$$f = g(M, P, V) \quad (2)$$

(where S is the Distinctiveness, M is the model, P is the speaker and V is the environment)

声纹安全理论分析：PROLE Score

- 声纹安全测量方法
  - 控制3个变量：$\boldsymbol{M}, \boldsymbol{P}$ and $\boldsymbol{V}$ (model, speaker and environment).

挑战1：减小模型、说话人和环境造成的偏差

基于**2**个大型数据集 → 评测超过**2600**个说话人 → 数据量超过**1000h** 语音 → 基于**3**个常用和商用模型

挑战2：构建超越语言约束的任意R, L, E和O的音频

超过1000小时的语料 → 切分成音素片段 → 重新组合为测试音频 → 超过**14,600,000**音频样本

一些有意思的结论

- 音素数量<10，**安全性随着音素数量增加而增加**
- 音素数量>10，安全性增加变平缓
- 当音素数量<5，**重复单词会增加安全性**
- 当音素数量>5，重复单词不会增加安全性
- 对不同模型，**元音普遍比辅音的安全性高**
- 存在一些magic音素，安全性特别高，一些特别低

声纹安全理论分析：PROLE Score

- **声纹安全测量结果**
  - 语音内**长度、丰富度、音素元素**影响声纹安全，**前后顺序无影响**
  - 唤醒词**增加语气词可以提升安全性**，Hey Google的安全性高于Google，类似中文的昵称，如小X小X。
  - 唤醒词**重复1-2次可明显提升安全性。**
  - 某些词语的**安全性较低，如包含音素[o]的词。**

声纹安全理论分析：PROLE Score

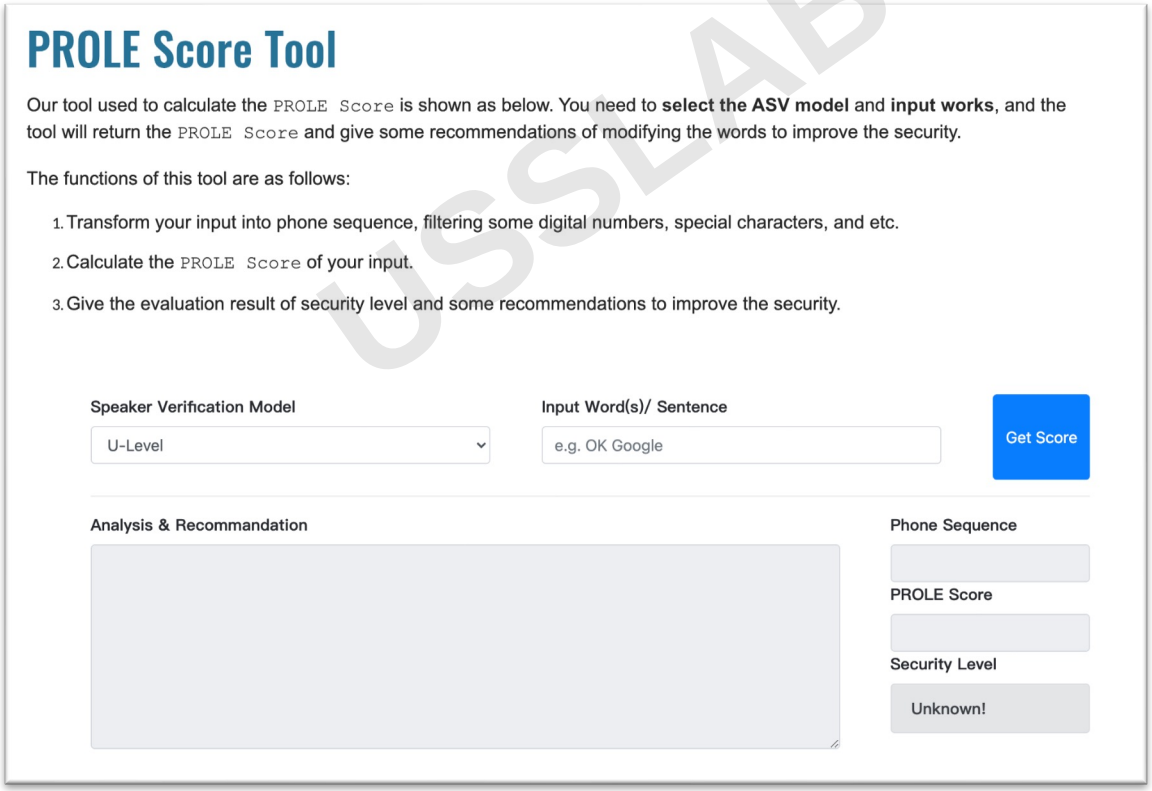
- 常用的语音助手唤醒词安全分析结果

English Wake-up Words Scores					Chinese Wake-up Words Scores				
Developer	Wake-up Words	i-v[1]	x-v[2]	U-L[3]	Developer	Wake-up Words	i-v	x-v	U-L
Amazon	Alexa	4.25	4.53	6.05	Alibaba	TianMaoJingLing	6.21	5.90	7.10
Amazon	Amazon	4.18	4.53	5.75	Baidu	XiaoDuXiaoDu	4.81	5.22	7.57
Amazon	Computer	5.40	5.34	6.32	Huawei	NiHaoXiaoE	4.90	5.18	7.08
Amazon	Echo	2.47	2.76	4.90	Huawei	NiHaoYoYo	4.42	4.91	6.12
Apple	Hey Siri	4.62	4.89	6.55	Huawei	XiaoEXiaoE	4.67	5.00	7.56
Google	Hey Google	4.77	4.04	5.75	JD	DingDongDingDong	5.64	5.39	7.46
Google	Ok Google	5.11	5.30	5.82	JD	Hey XiaoJingYu	5.66	5.81	7.34
Huawei	Hey Celia	4.56	4.56	6.44	Lenovo	NiHaoLianXiang	5.70	5.79	6.65
Microsoft	Hey Cortana	5.61	5.40	6.45	MeiZu	NiHaoMeiZu	5.11	5.33	6.64
Multiverse	Extreme	4.87	5.00	6.27	Microsoft	NiHaoXiaoNa	4.86	5.09	6.75
MyCroft	Hey Mycroft	5.47	5.51	6.21	Mobvoi	NiHaoWenWen	5.42	5.35	6.52
Nuance	Hello Dragon	5.76	5.82	6.07	OPPO	XiaoBuXiaoBu	4.63	5.07	6.98
OPPO	Hey Breeno	4.70	4.88	5.99	OPPO	XiaoOuXiaoOu	4.00	4.45	6.94
Samsung	Hey Bixby	5.10	5.24	6.09	Tencent	XiaoWeiXiaoWei	4.73	5.13	7.02
SoundHound	OK Hound	4.87	4.83	5.87	XiaoMi	XiaoAiTongXue	5.81	5.94	7.23

[1] Abbreviation for i-vector model. [2] Abbreviation for x-vector model. [3] Abbreviation for U-LEVEL model.

声纹安全理论分析：PROLE Score

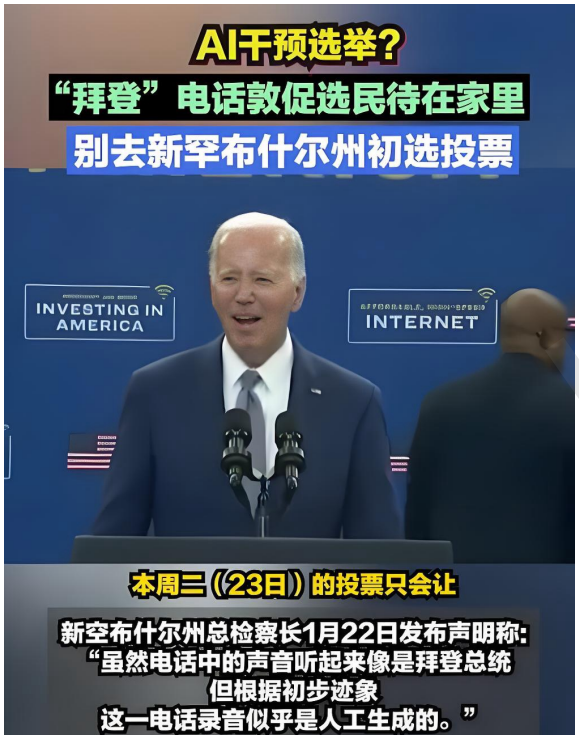
- 在线评分工具 PROLE Scoring
  - 输入: 自定义语音内容，选定声纹模型
  - 输出: 安全等级分数，安全提升建议



https://www.usslab.org/voiceprint-security/

延伸阅读：VoiceRadar合成语音检测

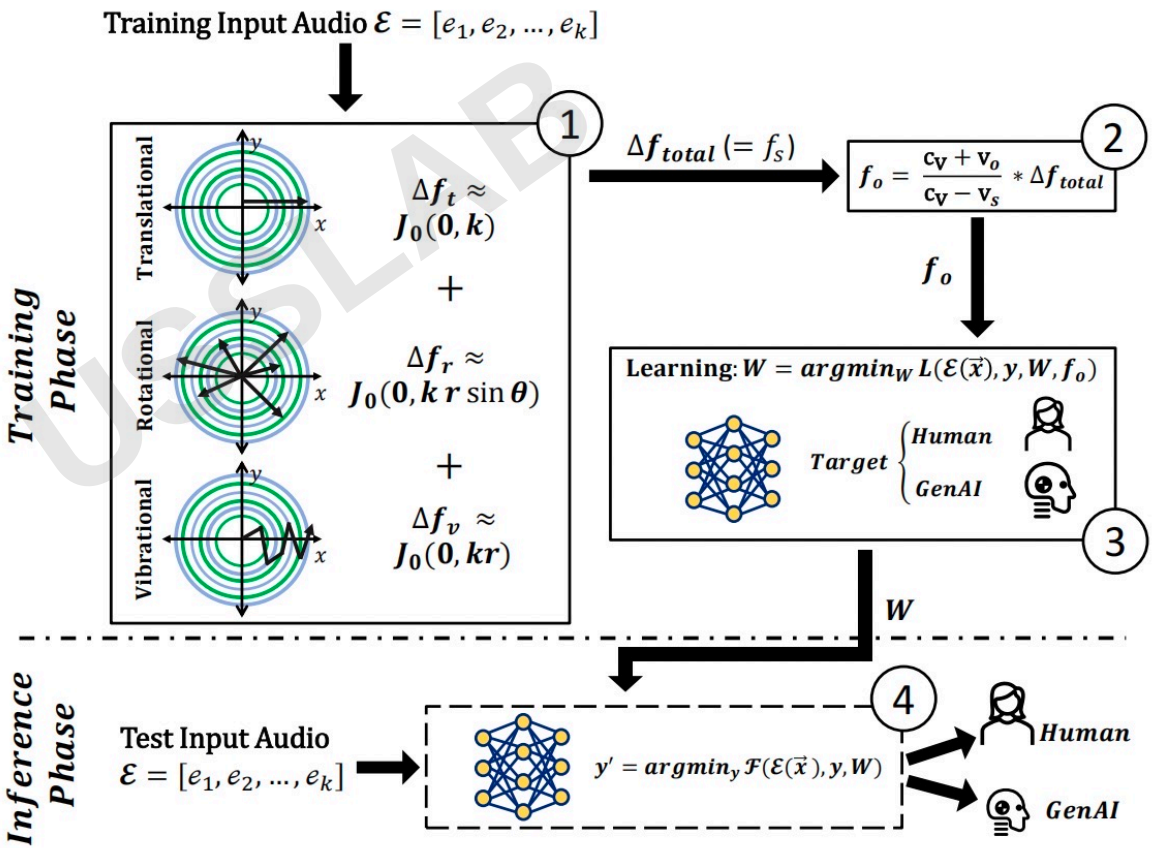
- 音频深度伪造检测威胁：
  - 场景：如何区分声音是真实的还是AI生成的？



AI合成语音检测 —— VoiceRadar

- **核心思想：利用物理世界的声学规律**

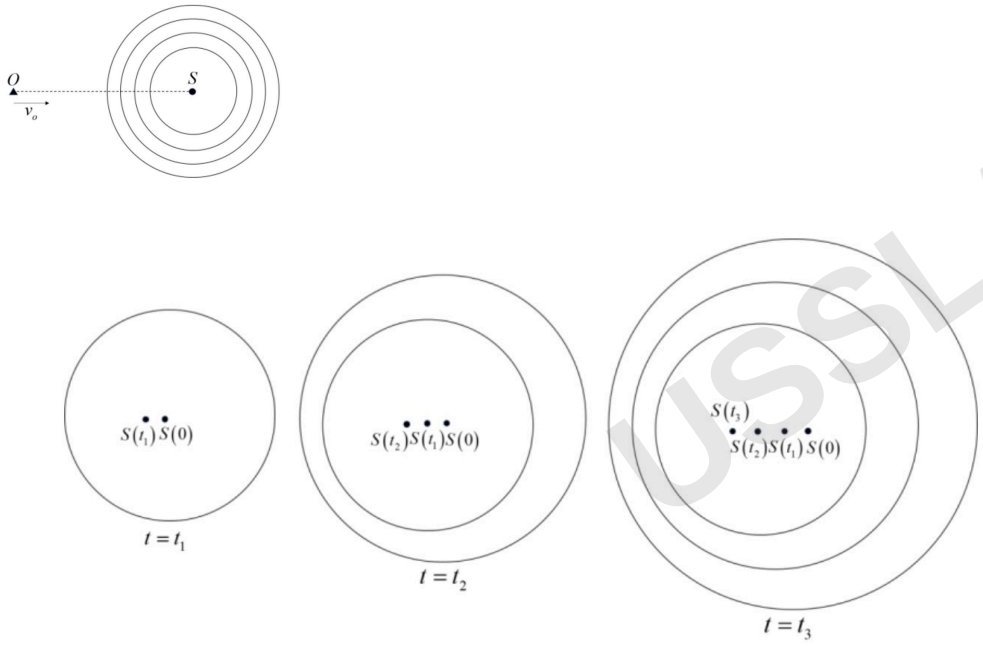
① 声波分析；
② 物理建模；
③ 训练基于物理特征的分类器；
④ 合成语音分类

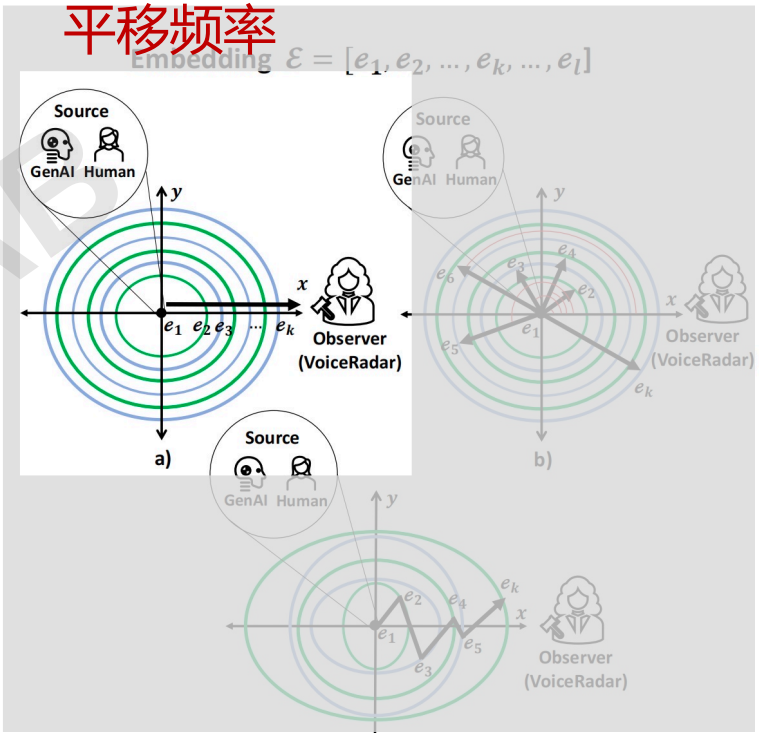


Kumari, Kavita, et al. “Voiceradar: Voice deepfake detection using micro-frequency and compositional analysis.” NDSS 2025.

AI合成语音检测 —— VoiceRadar

- **回归物理本质——多普勒效应**





平移频率：由于相对直线位移引起的频率变化

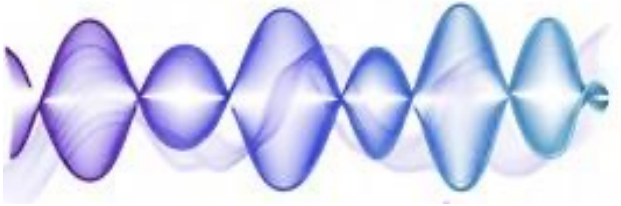
在音频中对应：声音信号最整体、最宏观的动向，一句话总体音高的缓慢上升或下降。

在VoiceRadar中：通过频率变化反推声源的“运动状态”

AI合成语音检测 —— VoiceRadar

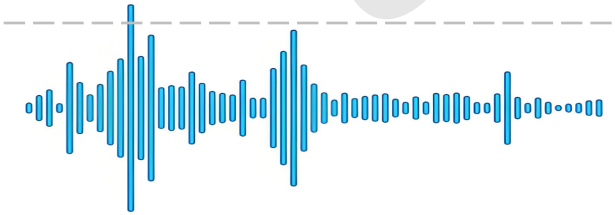
- **回归物理本质——鼓膜振动**

旋转频率

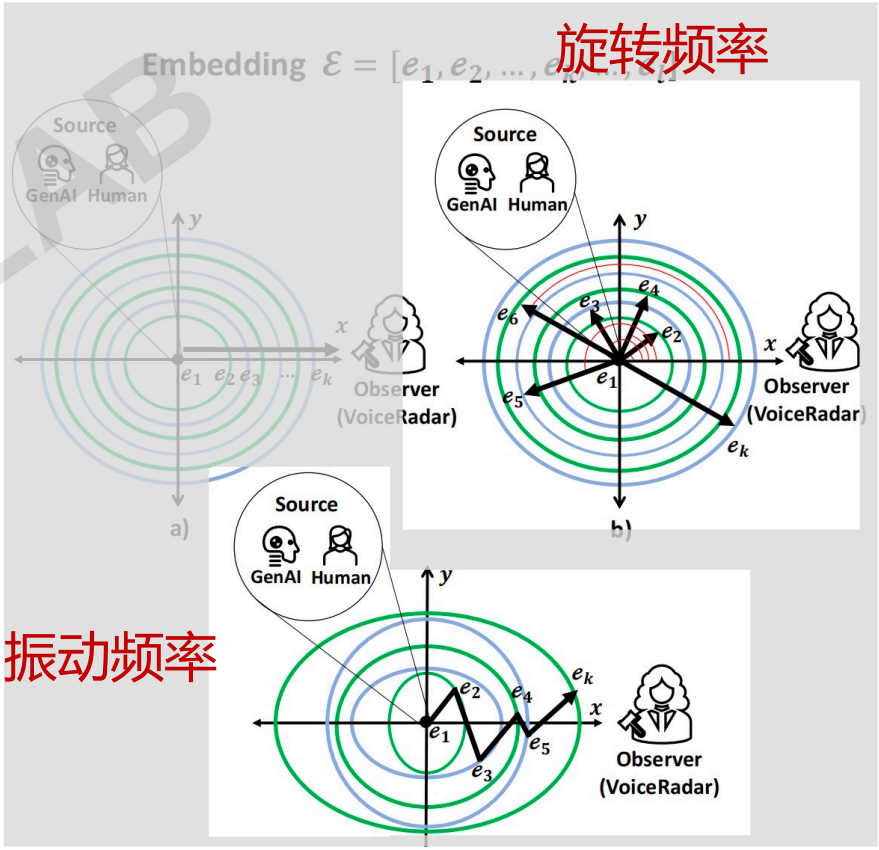


运动特点，绕轴旋转，移动时产生规律的、周期性的微小起伏，在音频中对应语调变化和音色调制，AI平铺直叙。

振动频率



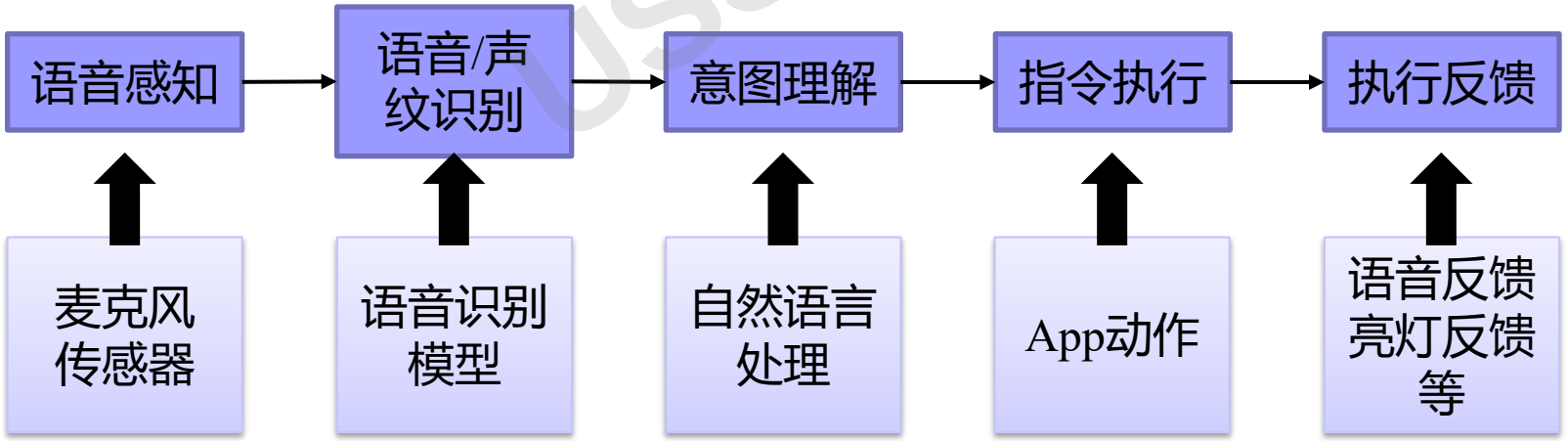
声音因为肌肉生理特性带有抖动而产生“颤音”或者听起来有点“模糊”、“毛糙”，AI生成声音不自然、机械规律的抖动。



4. 意图理解安全

回顾：智能语音系统工作流程

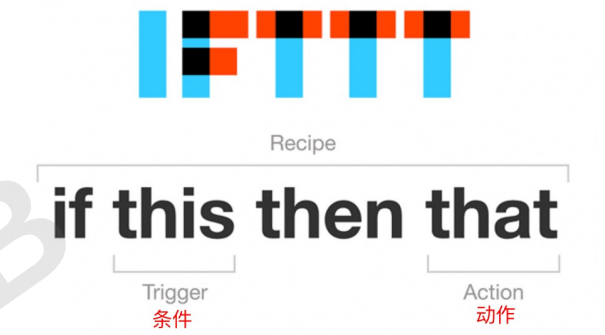
- **语音感知**：麦克风获取及后续处理电路，包括放大、滤波、AD转换
- **语音/声纹识别**：将语音信号转换为文本数据/识别出说话人身份
- **意图理解**：将文本数据转换为可理解的指令内容
- **指令执行**：执行指令内容
- **执行反馈**：通过声音或者执行动作进行反馈



举例：通过Amazon Echo打开窗帘……

意图理解——IFTTT

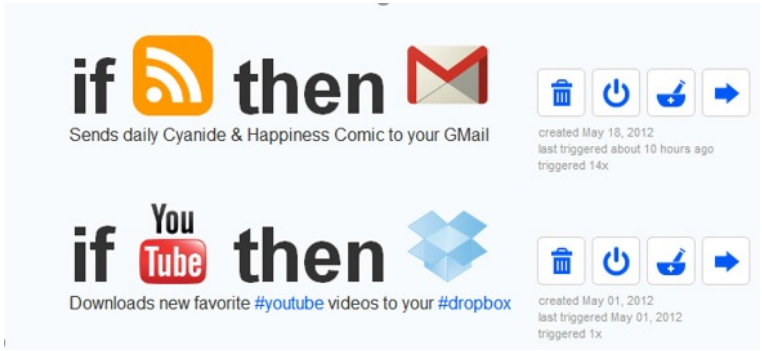
- 智能语音系统按照识别的指令，通过自带的或者第三方APP调用执行机构比如电机进行动作执行
  - 如：天气问答、门窗控制

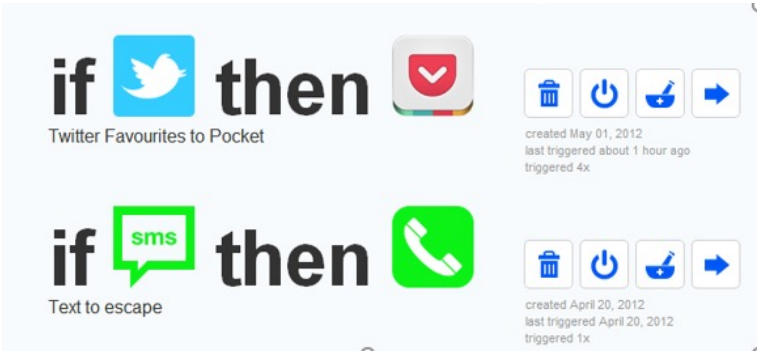


- **第三方APP案例：IFTTT**

IFTTT是“if this then that”缩写，旨在帮助人们利用各网站的开放API，网站或应用衔接，完成任务。

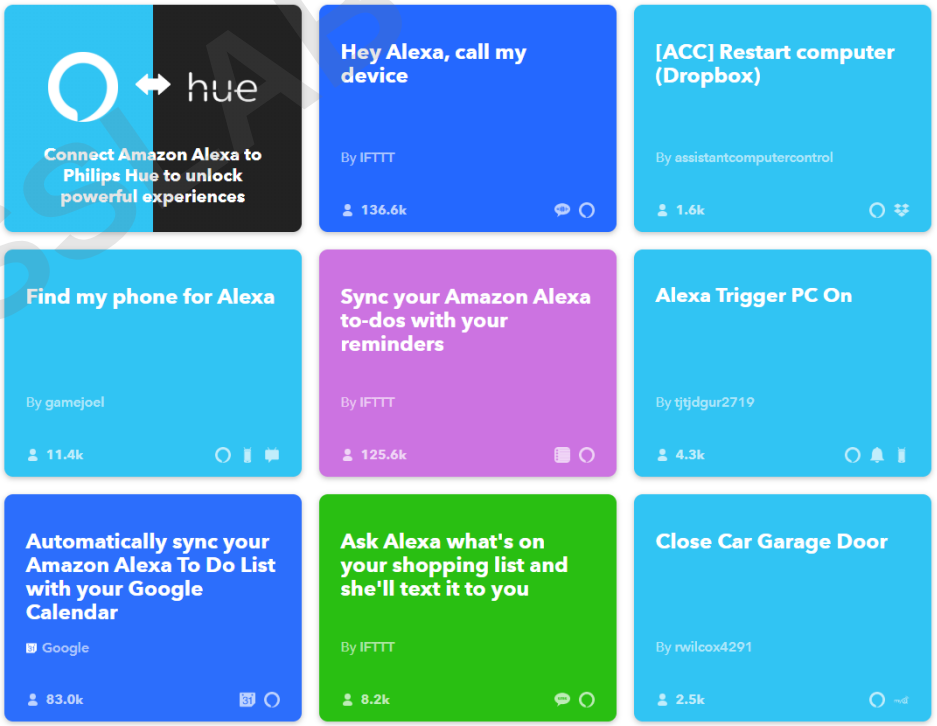
IFTTT通过流程将各种信息串联起来，然后再集中把信息呈现，解决信息的冗杂，收取或关注重要信息的问题





意图理解——IFTTT

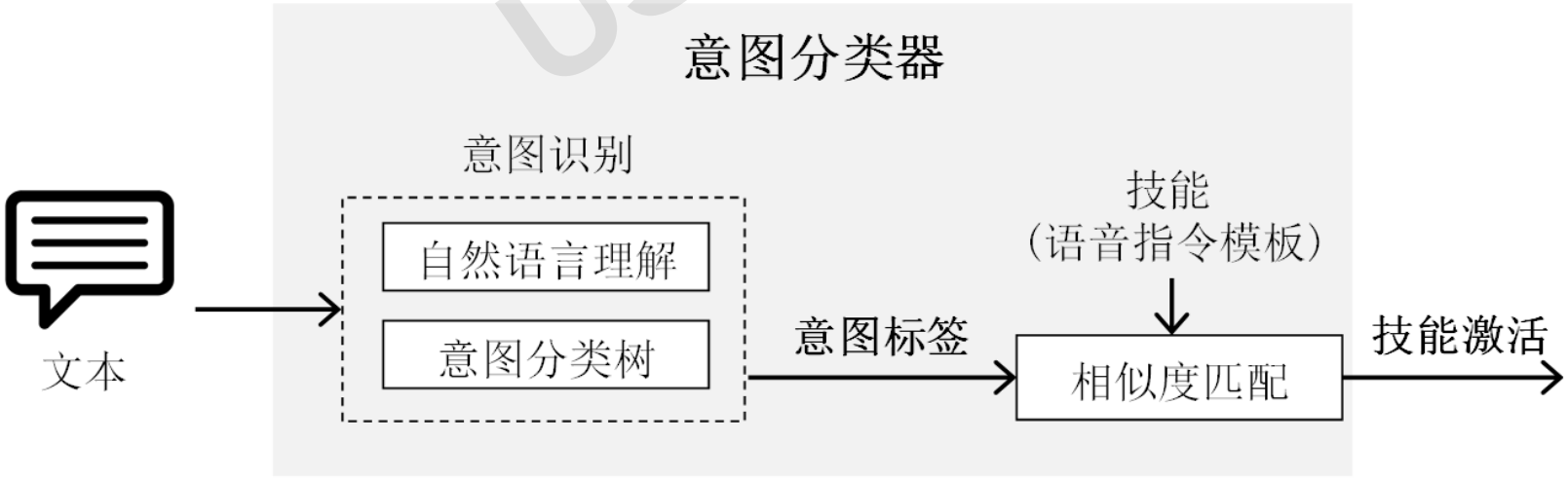
- Amazon Alexa：安装IFTTT到Alexa中，使用IFTTT及其使用第三方库，通过skill的方式，让语音实现各类操作
  - 开关智能门锁
  - 开关灯
  - 呼叫设备
  - 找寻手机
  - 购物清单查询
  - 重启电脑
  - ……



- **Skill关键：意图理解**

意图理解安全

- **意图理解**：用于识别和理解语音输入的意图或目的，包括执行动作、传递信息等
- 由意图分类器(Intent Classifier)实现
- **意图分类**：识别文本的意图标签并给出置信度分数，并将标签和各类语音应用技能相匹配，用于激活技能



意图理解安全——自然语言处理

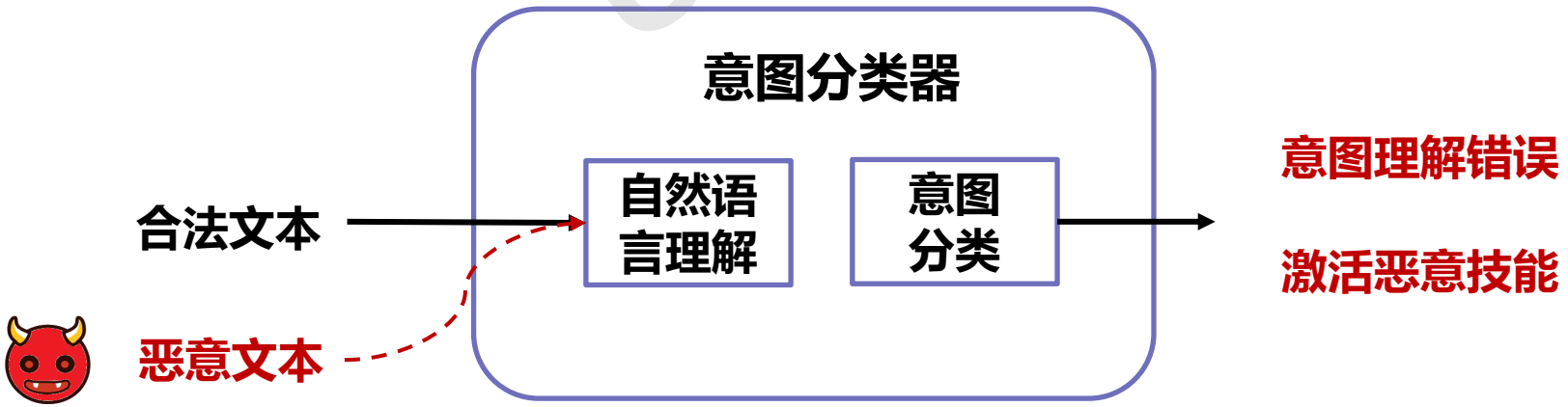
- **自然语言：**自然语言是人类日常交流中使用的语言，即“讲人话”
  - 自然语言：“我背有点驼”
  - 非自然语言：“我的背部呈弯曲状”

- **自然语言理解存在歧义，会增加意图理解安全风险，包括：**
  - **词汇歧义**：一个词多种含义，例如苹果可以指公司也可以指水果
  - **句法歧义**：由于句子结构导致多种解释，例如“看见山上的人很高兴”
  - **语义歧义**：句子的语义导致的歧义，词与词之间的关系或者整个句子的含义不明确，例如：“校长说衣服上除了校徽别别别的”、“滴滴拉拉布拉多取决于拉布拉多拉的多不多”
  - **语境歧义**：同一个词或句子在不同的语境中可能有不同的含义

- **攻击者可以利用自然语言歧义特性构造攻击语句实现意图理解攻击**

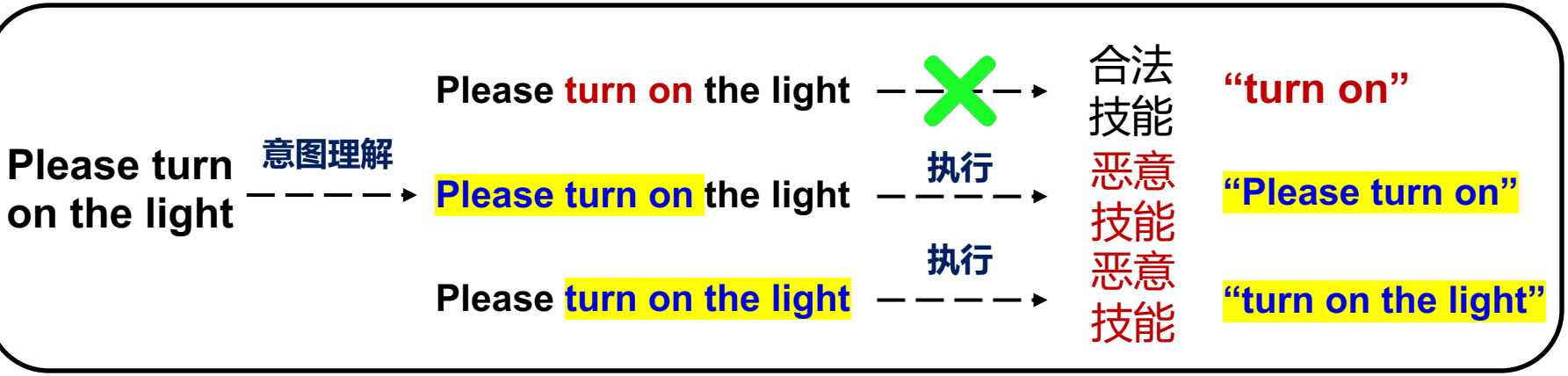
意图理解安全——意图分类

- **意图理解攻击**
  - **攻击原理**：攻击者利用用户意图理解环节算法及逻辑漏洞，通过篡改用户意图理解结果，并上传恶意技能，诱使用户合法意图被理解为其他意图，并激活恶意技能的一类攻击
  - **攻击方法**：恶意技能构造
  - **攻击结果**：窃取隐私信息、执行错误操作、操控语音设备



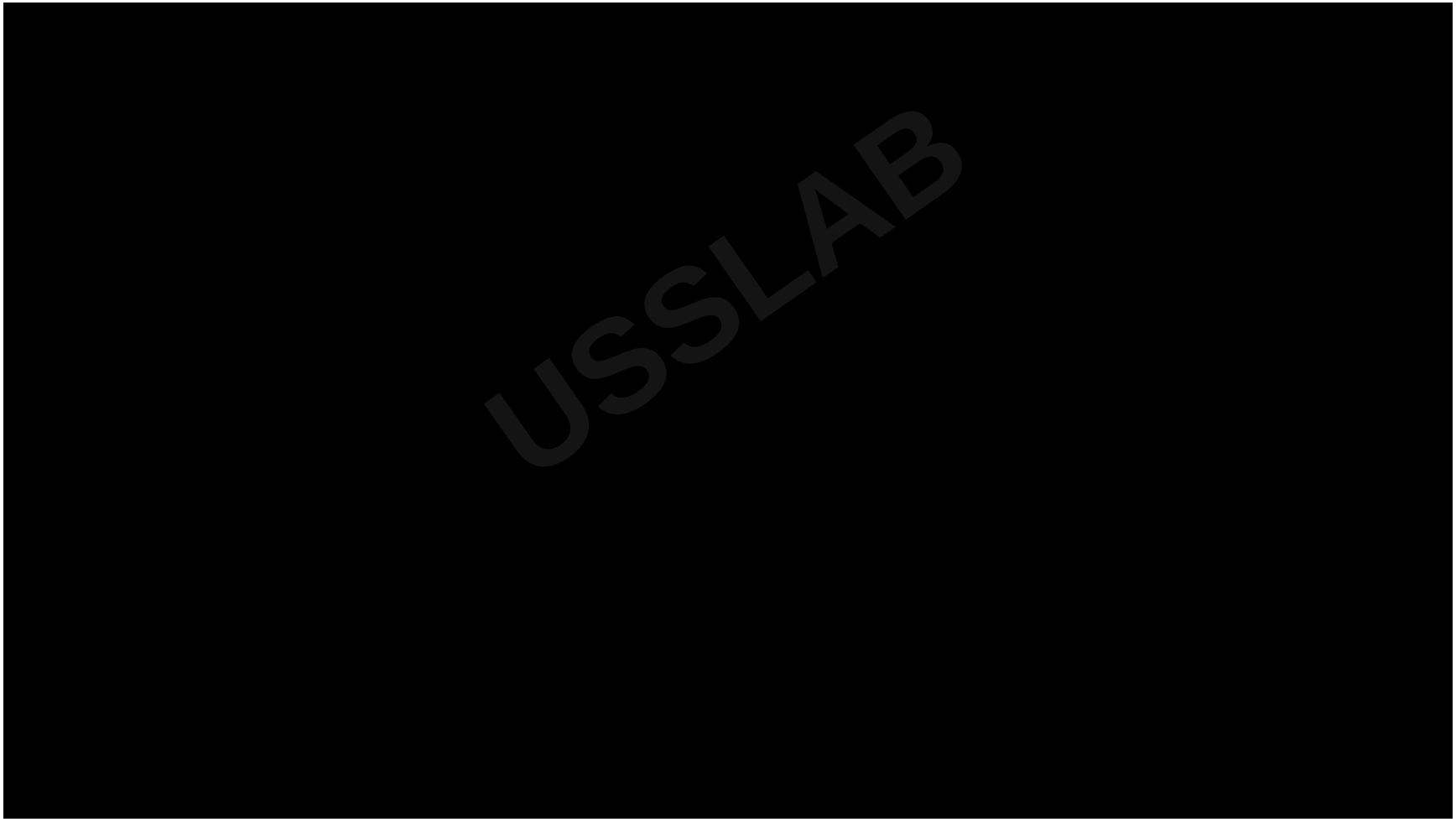
意图理解安全——恶意技能构造

- **方法1：同音异义词**
  - 利用**发音相同但含义不同**的词触发恶意语音技能，例如亚马逊的Alexa音箱会混淆“wet”和“what”、“lung”和“lang”等
- **方法2：目标词变体**
  - 利用用户说话时的特定**口语变化触发攻击技能**，包括添加语气词、敬语。
  - 自然语言理解倾向于把**语音与指令模板较长的技能进行匹配**



案例：Skill Squatting Attack

Skill: Rat game → rap game



Kumar D, Paccagnella R, Murley P, et al. Skill squatting attacks on Amazon Alexa, USENIX Security'18

用户语音隐私安全

- **语音内容隐私**：语音系统在收集语音数据的过程中，采集或上传了用户未授权的语音数据

- **关联信息隐私**：语音数据被获取之后，攻击者可提取、推断出与语音数据关联的用户个人信息用于其他攻击场景

- **数据共享隐私**：语音系统在未经用户允许的情况下将收集到的用户语音数据与第三方共享造成的用户隐私泄露问题

- **语音生态隐私**：由政府监控或语音系统服务商跟踪用户语音数据而造成的用户隐私问题

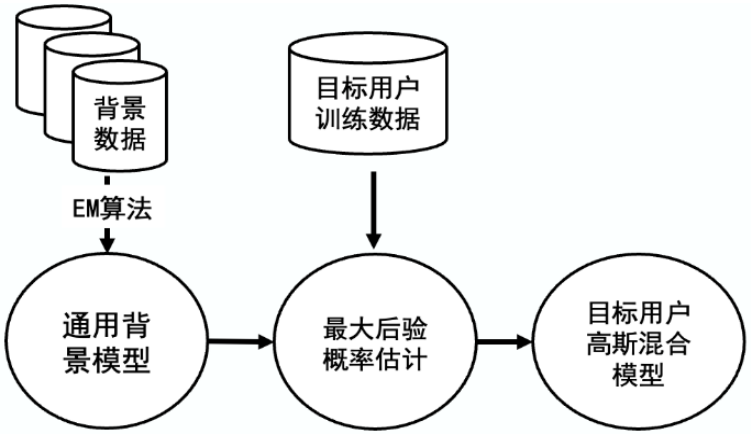
本章总结

- 了解语音基本知识
- 掌握物联网语音安全的含义
- 理解智能语音系统工作流程
- 掌握针对物联网语音安全四种攻击，了解定义、原理并能分析基本案例
  - **信号感知安全**
  - **内容识别安全**
  - **声纹识别安全**
  - **意图理解安全**

BACKUP SLIDES

基础知识：声纹识别模型GMM-UBM

- 高斯混合模型 (GMM)：**一系列的高斯分布函数的线性组合，理论上GMM可以拟合出任意类型的分布**
- 充分训练的GMM模型，可以较好地表征说话人空间。
  - 训练阶段：
    - 用大量不同说话人的数据训练**UBM (Universal Background Model)** 模型 (本质是多个GMM模型)
    - 针对每个说话人，在UBM模型基础上做**MAP (Maximum a Posteriori Probability)**，得到每个说话人的speaker模型



基础知识：声纹识别模型GMM-UBM

- 高斯混合模型 (GMM)：**一系列的高斯分布函数的线性组合，理论上GMM可以拟合出任意类型的分布。**
- 充分训练的GMM模型，可以较好地表征说话人空间。
  - 测试/推理阶段：对于某条语音Y，定义如下：
    - H_0：Y是说话人S说的
    - H_1：Y不是说话人S说的
    - 分别在speaker模型和UBM模型上计算似然度：

$$\text{Score} = \frac{\text{Y}|\text{H}_0}{\text{Y}|\text{H}_1} \begin{cases} \geq \theta, accept\ H_0 \\ < \theta, reject\ H_0 \end{cases}$$

基础知识：声纹识别模型I-Vector

- 说话人矢量因子（Identity-Vector，I-Vector）为了解决GMM-UBM中信道干扰问题，并进一步**降低计算复杂度和降低语料数据量**的算法。
- I-Vector**将语音的信道因子和说话人因子合二为一**，建模为**总体因子**。因此**I-Vector即包含了说话人之间差异又包含信道间差异**，其公式定义如下：

$$s = m + T\omega$$

***s* (super-vector):** 由给定说话人的语音计算高斯均值超矢量（GMM建模用户）

***m* (UBM's mean super-vector):** 通用背景模型的高斯均值超矢量

***T* (total-variability matrix):** 全局差异空间矩阵

***ω* (I-Vector):** 全局差异空间因子

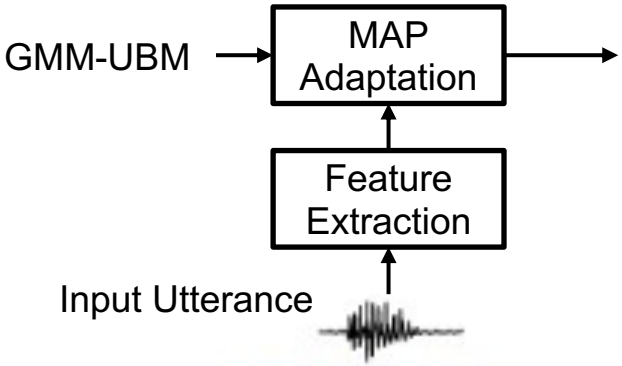
I-Vector也可以视为一种被提取得到的说话人特征

基础知识：声纹识别模型I-Vector

- I-Vector求解过程

$$s = m + T\omega$$

  - 训练GMM-UBM通用背景模型，根据训练语料库的所有音频得到高斯均值超矢量s
  - 基于EM算法，迭代估计全局差异空间T矩阵
    - E-Step：计算说话人i对应隐变量ω_i的后验概率分布，即该说话人第n段语音的后验均值$\omega_{i,h}$以及后验相关性矩阵的期望。
    - M-Step：基于最大似然值重新估计，更新T矩阵 (一般为10次迭代)
  - 任意音频的I-Vector提取：



$$s = \begin{Bmatrix} s_1 \\ s_2 \\ \cdots \\ s_n \end{Bmatrix} = m + T\omega$$

由于待提取音频的高斯均值矢量已知，T矩阵已知，则可以获得I-Vector：$\boldsymbol{\omega}$