Transition Page

专题：音频功率放大器

科普 | 胆机、石机是什么？音频功率放大器类型剖析

Think 周鸿祎 音响 多少钱？有几套？


周鸿祎（360公司创始人）

声音转换成特征以后……

声音的研究与应用

深度学习及应用

声音转换成特征以后……

可选作业：语音类工具复现

声音处理的应用

会议系统

降噪效果和实时处理性能

“高质量的降噪效果和实时处理性能是视频会议中音频降噪的两大挑战

AI学习心脏病喘息声

AI学习心脏病发时特有的喘息声

- 相关研究也以论文的形式发表在了Nature上
  - 有兴趣的朋友们可以自行前往细读：
  - https://www.nature.com/articles/s41746-019-0128-7

Article | Open Access | Published: 19 June 2019

Contactless cardiac arrest detection using smart devices

Justin Chan, Thomas Rea, Shyamnath Gollakota & Jacob E. Sunshine

npj Digital Medicine **2**, Article number: 52 (2019) | Cite this article

18k Accesses | **20** Citations | **713** Altmetric | Metrics

AI学习心脏病发时特有的喘息声

- 研究人员对此并非只是示范可行性，事实上，他们已经成立了一间名为Sound Life Sciences的公司，准备让这个**技术商品化**。
  - 要完美触发这个系统的前提是，收音设备必须持续打开，而如果直接套用在现有的Alexa或Google Assistant上，难免会有隐私上的疑虑。
  - 比较可能的做法，是将它的判读系统独立建置在机器上，不用通过云端，但这么一来就会变成打造专门的机器，而不是利用愈来愈普及的智能喇叭了。
  - 医疗器械：同学——深圳

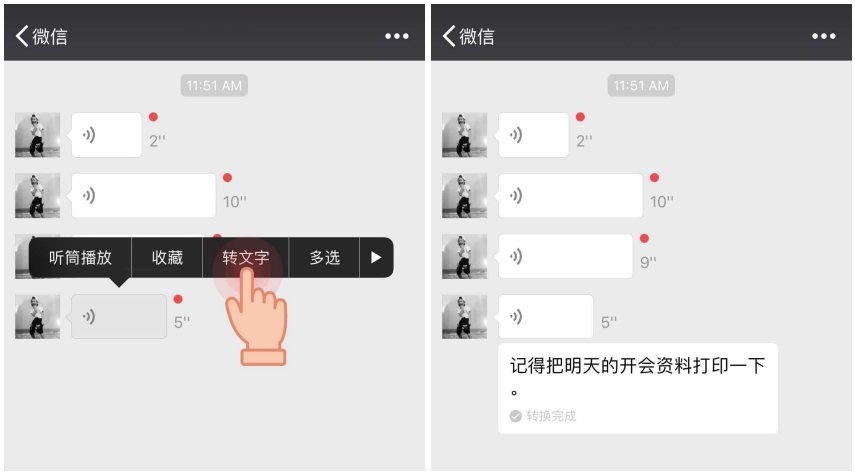
关于声音的其它研究领域

领域/产品/技术/实践

领域1 语音识别/声音生成

语音识别


拥有1600+生活技能
你的生活，交给小爱打理
“小爱同学，今天天气怎么样？”
“小爱同学，距离七夕节还有几天？”
08: 16
xiaomi



微软工程师如何让语音识别准确度达到人类同等水平

- 微软人工智能和研究所一篇论文：
  - 报告中提到了一个语音识别系统，这个系统与专业的打字员相比有过之而无不及。
  - 研究人员在报道中讲到，这个系统的单词错误率(WER)为5.9%，就在上个月的报道中WER还是6.3%。
- 程碑意味：
  - 史上第一次，一个计算机可以像一个自然人一样识别会话环境中的单词。

讯飞听见
所 听 即 所 见

语音识别技术里程碑：错误率降至5.1%，超过专业速记员

- 微软语音和对话研究团队负责人黄学东宣布微软语音识别**错误率由5.9%进一步降低到5.1%**

- 语音研究领域仍然挑战重重：
  - 例如嘈杂环境、录音距离较远场景下的语音识别，方言识别，有限训练数据条件下的语音识别或较少人使用的语言的语音识别，这些距离达到人类相近水平还相差甚远。
  - 引入了**CNN-BLSTM**(convolutional neural network combined with bidirectional long-short-term memory，**带有双向LSTM的卷积神经网络**)模型
  - 2017年8月20日

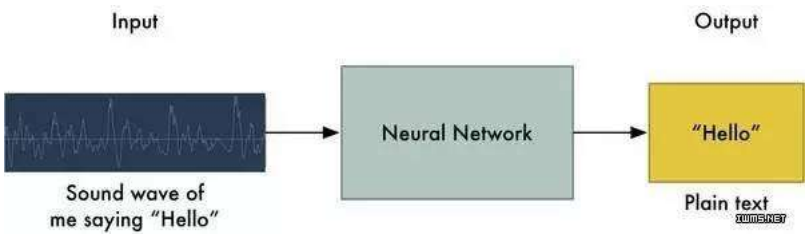
Microsoft Research
微软最前沿的科技信息

深度学习进行语音识别

如何利用深度学习进行语音识别

机器学习并不总是黑箱

- 深度学习进行语音识别的**核心**:
  - 将声音录音输入神经网络，然后训练神经网络来生成文本 。



- 最大的问题是语音会随着速度变化。

Think 语音会随着速度变化?

一个人可能很快地说出「Hello！」，而另外一个人可能会很缓慢说「heeeellllllllllllooooo!」。这就产生了一个更长的声音文件和更多的数据。

这两个声音文件本应该被识别为完全相同的文本「hello！」而事实证明，把各种长度的音频文件自动对齐到一个固定长度的文本是很难的一件事情。

TTS

TTS(文-语转换)系统

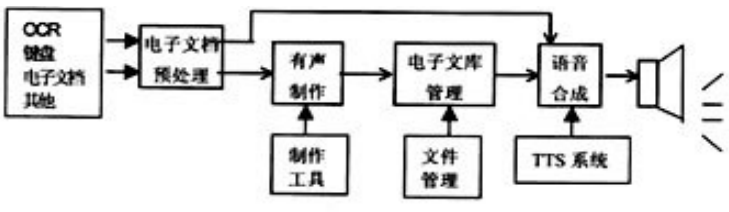
- 世界强国竞相研究的热点之一。
  - 世界上已研究出多种语言的TTS系统，如汉、英、法、日、德等。
  - TTS系统最根本的问题便在于它的自然度。

语音合成技术

TTS的典型应用

声音知识

- TTS已应用到信息咨询、电话银行、办公自动化等各个方面，它把声音和文字、图像集成在一起，增强了人们的理解和阅读兴趣，使人与计算机之间的交流变得“亲切”和“友好”。



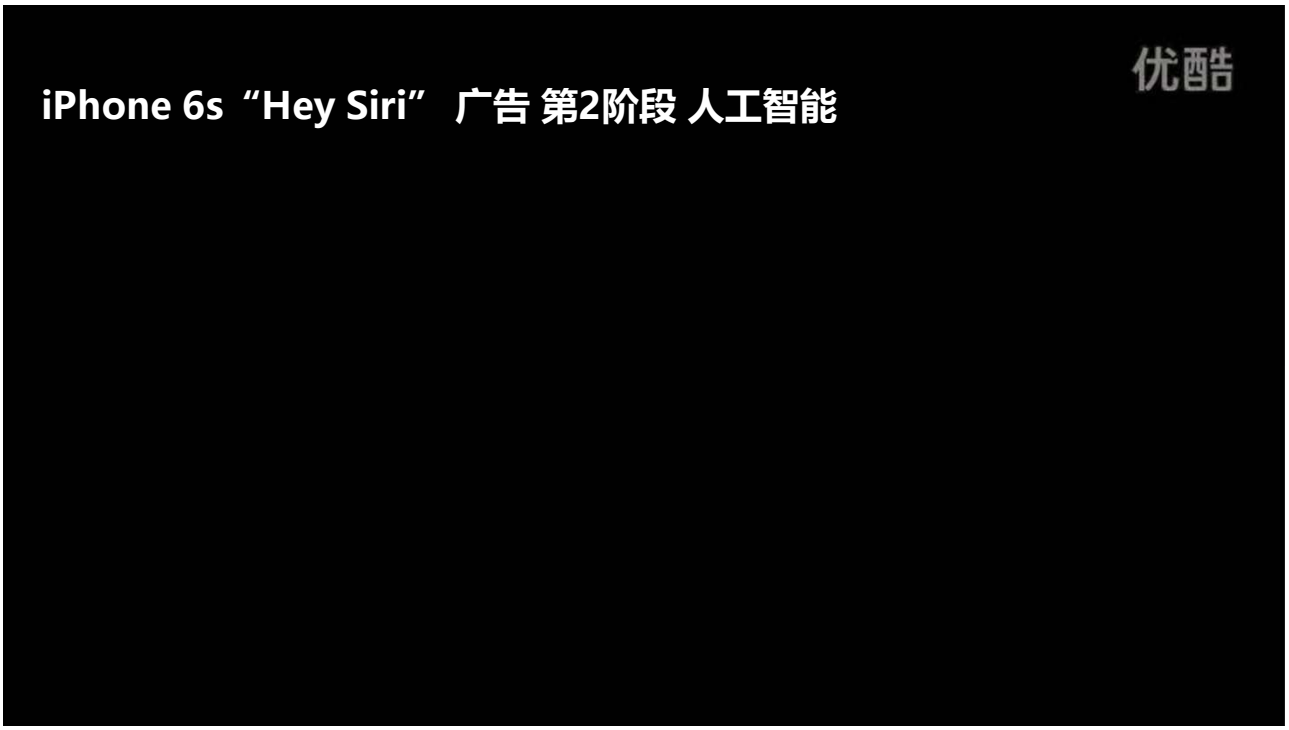
电子文档有声输出的原理框图

可选作业：TTS的应用

TTS的调用和演示等

基于深度学习的语音生成问题

- **Siri 的 TTS 系统**的目标是训练一个基于深度学习的统一模型，该模型能自动准确地预测数据库中单元的目标成本和拼接成本。
  - 通过语音增强可以有效抑制各种干扰信号，增强目标语音信号；
  - 有效的语音增强算法：
  - 一方面可以提高语音可懂度和话音质量，
  - 一方面有助于提高语音识别和声纹识别的鲁棒性。
- 经典的语音增强方法包括谱减法、维纳滤波法、最小均方误差法，上述方法基于一些数学假设，在真实环境下难以有效抑制非平稳噪声的干扰。



语音识别技术

科大讯飞AI 语音合成/识别

翻译机

语音交互的重要产品

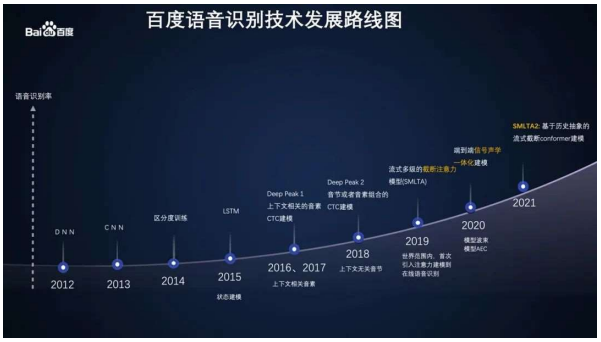
语音助手

语音助手是一款智能型的应用，通过智能对话与即时问答的智能交互，实现帮助用户解决问题，其主要是帮忙用户解决生活类问题。

Think 微软 Cortana vs 苹果 Siri?

开源的语音工具

百度Deep Voice 3：全卷积注意力机制TTS系统



- 2017年 2 月份，百度提出了完全由深度神经网络构建的高质量文本转语音（TTS）系统 Deep Voice。
  - 2017年5月：推出了第二个版本。
  - 2017年10月：发布了 Deep Voice 3，该研究的论文已经提交 ICLR 2018 大会。

一行代码轻松搞定中英文语音识别、合成、翻译核心功能

- GitHub 地址：
  - 百度语音识别系统DeepSpeech+嘈杂环境识别率超谷歌苹果(报道)
  - https://github.com/PaddlePaddle/PaddleSpeech

语音识别模块种类	数据集	模型种类	链接
语音识别	Aishell	DeepSpeech2 RNN + Conv based Models	deepspeech2-aishell
		Transformer based Attention Models	u2.transformer.conformer-aishell
	Librispeech	Transformer based Attention Models	deepspeech2-librispeech / transformer.conformer.u2-librispeech / transformer.conformer.u2-kaldi-librispeech
对齐	THCHS30	MFA	mfa-thchs30
语言模型	Ngram 语言模型		kenlm
	TIMIT	Unified Streaming & Non-streaming Two-pass	u2-timit
语音翻译（英译中）	TED En-Zh	Transformer + ASR MTL	transformer-ted
		FAT + Transformer + ASR MTL	fat-st-ted

语音识别包含声学模型和语言模型

语音合成模块类型	模型种类	数据集	链接
文本前端			tn / g2p
声学模型	Tacotron2	LJSpeech	tacotron2-ljspeech
	Transformer TTS		transformer-ljspeech
	SpeedySpeech	CSMSC	speedyspeech-csmsc
	FastSpeech2	AISHELL-3 / VCTK / LJSpeech / CSMSC	fastspeech2-aishell3 / fastspeech2-vctk / fastspeech2-ljspeech / fastspeech2-csmsc
声码器	WaveFlow	LJSpeech	waveflow-ljspeech
	Parallel WaveGAN	LJSpeech / VCTK / CSMSC	PWGAN-ljspeech / PWGAN-vctk / PWGAN-csmsc
	Multi Band MelGAN	CSMSC	Multi Band MelGAN-csmsc
声音克隆	GE2E	Librispeech, etc.	ge2e
	GE2E + Tactron2	AISHELL-3	ge2e-tactron2-aishell3
	GE2E + FastSpeech2	AISHELL-3	ge2e-fastspeech2-aishell3

语音合成主要包含三个模块：文本前端、声学模型和声码器

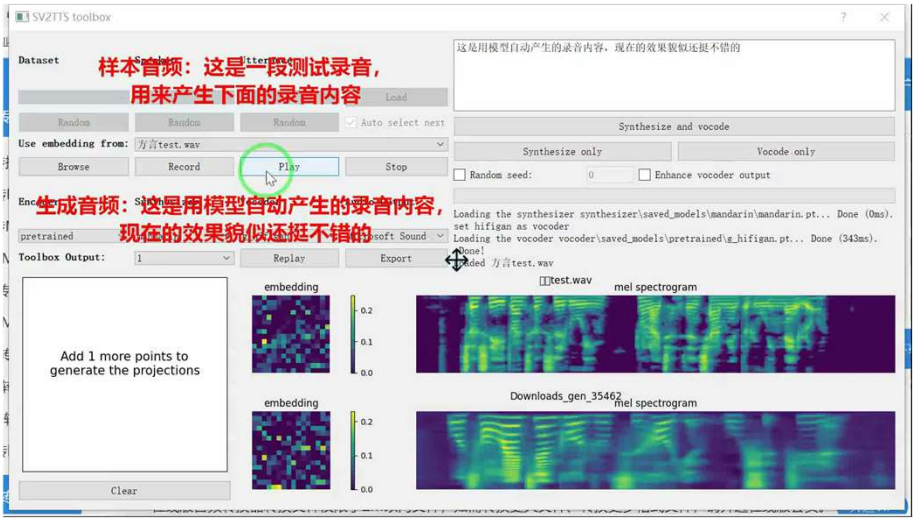
领域3 声音的“生成”

克隆声音

5秒就能“克隆”本人语音!

- GitHub博主Vega最新的语音克隆项目MockingBird，能够在5秒之内克隆任意中文语音，并用这一音色合成新的说话内容。
  - 项目地址：
  - https://github.com/babysor/MockingBird/blob/main/README-CN.md
  - 训练者教程：
  - https://vaj2fgg8yn.feishu.cn/docs/doccn7kAbr3SJz0KM0SIDJ0Xnhd

“克隆”本人语音 声音克隆.wmv



假声音/变声

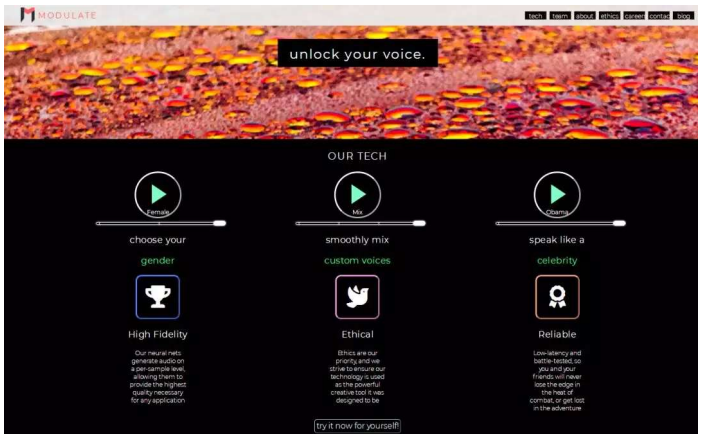
“假声音”也来了，手把手教你造一只柯南的蝴蝶结变声器

- 变音技术：江湖上确实流传了几种，不过加持了机器学习和深度学习，这种技术不再是简单的语音滤波器。
  - https://modulate.ai/
  - 一段是合成的语音。

  - https://mp.weixin.qq.com/s/J7fs1-2FUoz9p_24RtOTaw

网站地址：https://modulate.ai/

- 创建这个网站的三个小伙伴，有两个来自麻省理工，还有一个来自加州大学洛杉矶分校。
  - 对于游客，这个网站给出了几个适用的声音，对于想定制名人声音的用户，还得通过官网给出的联系方式联系他们。
  - 据网站介绍，合成的声音是是采用神经网络训练来训练，具有低延迟性以及实时性。

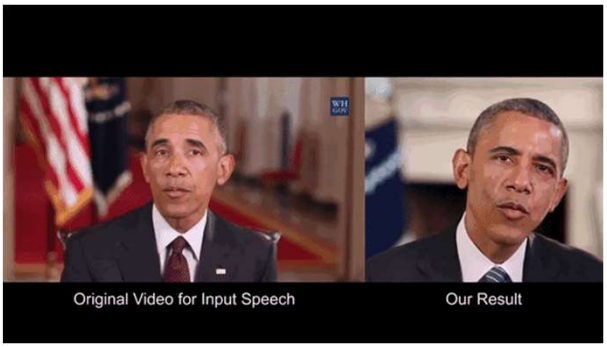


这时候再加上“假脸”？

靠换脸技术”出演”《射雕英雄传》的杨幂

假脸+声音

– “假脸”技术大肆盛行，与之配套的“假声音”上线后，更能生成无缝衔接的假视频，让假戏做足，真假难辨。



领域4 虚拟主播/歌手

虚拟主播

央视AI手语主播——通过朱广权魔鬼面试

- 手语主播并非真人，而是一名来自百度智能云的虚拟数字人。
  - 2022年已正式上岗冰雪盛会，将在各类冰雪赛事中，为2780万听障人士提供24小时不间断的手语服务。
  - 根据测评，其手语可懂度能达到85%以上，与主流的中英、中日机器翻译结果相差无几。

AI虚拟数字人直播

AI配音

GAN生成音乐

谷歌大脑团队用GAN生成高保真音乐

- GAN生成高保真音乐
  - 谷歌 AI 总统帅 Jeff Dean 也被这个研究吸引，大加赞赏，并建议大家试听一下更多样本音乐。
  - 谷歌大脑团队最新ICLR论文提出用GAN生成高保真音乐的新方法，速度比以前的标准WaveNet快5万倍，且音乐质量更好！

虚拟歌手

国内团队研究成果

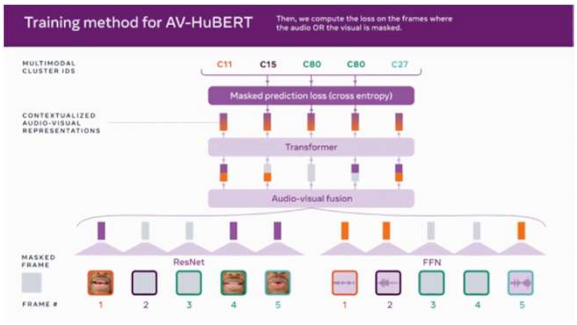
机器人模仿-歌手

领域5 多模态/物理领域

多模态应用

AI +唇语

AI 看唇语，在嘈杂场景的语音识别准确率高达75%



- 为了研究视觉效果，尤其是嘴部动作的镜头，是否可以提高语音识别系统的性能。
  - Meta 的研究人员开发了 Audio-Visual Hidden Unit BERT (AV-HuBERT)，这是一个通过观看学习和听人们说话来理解语言的框架。
  - AV-HuBERT 也是多模态

AI+图像处理->盲人?

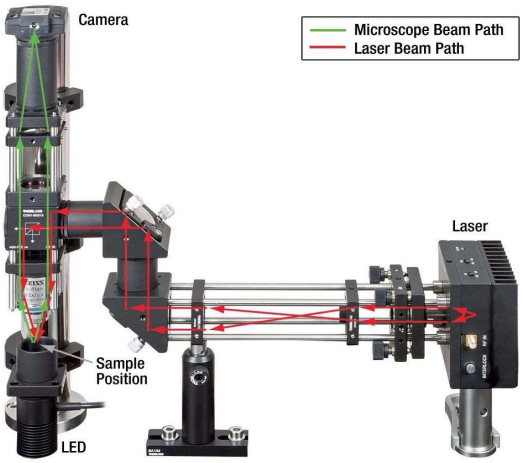
让他们听见世界：用多模态预训练模型，铺设数字化“盲道”

隔空取物

阿什金在90多岁高龄获得了2018年诺贝尔物理学奖

- 2018年的诺贝尔物理学奖颁给了光镊技术
  - 利用光镊可以操控许多微小物体，但是光镊却有一个缺点，那就是仍然保留了光的性质不能穿过非透明物质。
  - 但是最新的研究——声镊技术，可以完全克服这一缺点，可以说是诺奖的2.0版本。

注：1986年，阿什金等人指出将单束激光高度聚焦，在激光束焦点处可以将微粒稳定地捕获，这种技术被称为光镊技术。



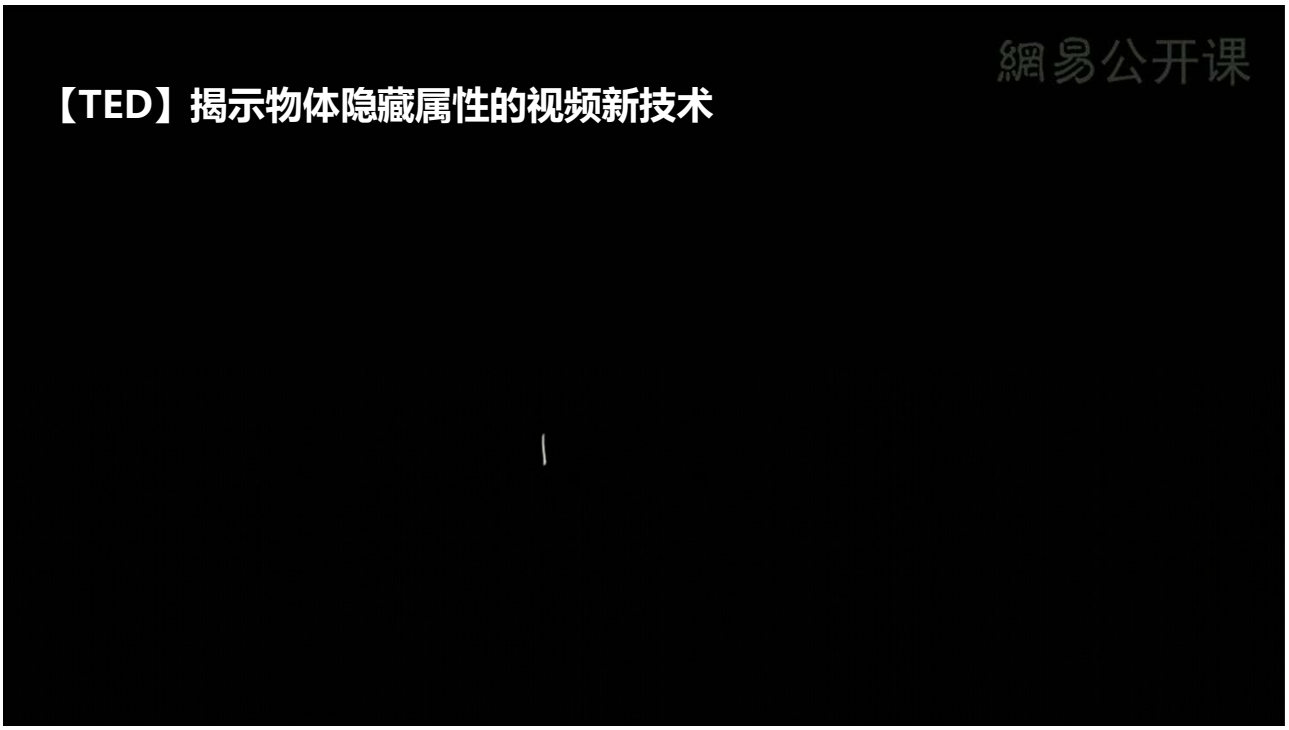
一台光镊仪器。图片来源：thorlabs

Think 声音如何存储?

奇妙的存储方式?

领域6 声音的存储

新的存储或获取方式



领域7 身份认证