CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition

# Demo地址：

<https://sites.google.com/view/commandersong/>

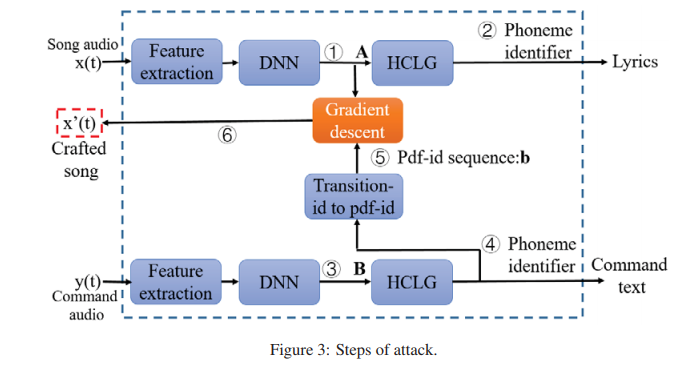
可以在现实中把语音识别系统给欺骗，具体来说他们往这些音乐片段里面添加一些Command的音素，然后把这些音素合成到目标音乐片段中，然后人耳听起来没有什么异常，但是语音识别系统会把音乐里面嵌入的指令识别出来。

# Motivation

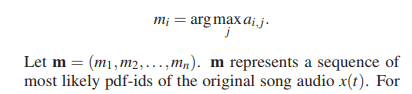
语音识别领域的对抗样本研究较少，有一篇工作是攻击的GMM的语言识别系统但是它是通过添加噪声的方式然后被人耳识别出来，另外一篇工作在人耳不可听到的频率里面添加语音指令。

# 方法

作者首先找到了一个开源的语言识别系统Kaldi，然后基于这个系统进行白盒攻击，寻找对抗样本。

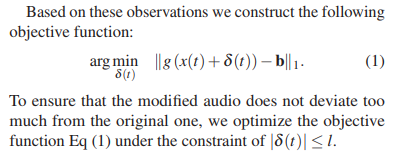


假设原歌曲为,目标指令音频为，把两段音频输入到Kaldi中，得到原音乐的音素矩阵A，和目标音乐的矩阵B。A包含了每一帧的各个音素的概率分布，假设有n个frame,k个音素，表示第i个frame,第j个音素的概率。然后使用贪心策略，得到各个frame上的最大概率音素：



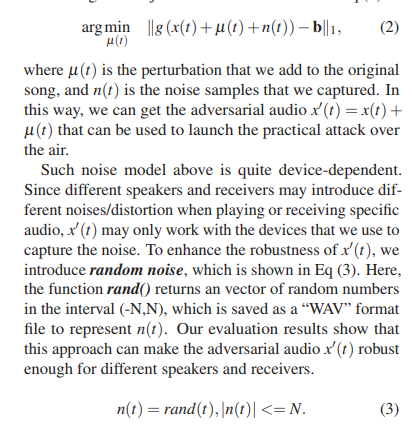
m就是各个frame的音素数列。 表示把音频x的音素数列提取出来。

同样的，也提取出目标指令音频的音素数列b。接下来，往x(t)加入扰动使得， 距离最小化。然后使用梯度下降来完成这个工作。



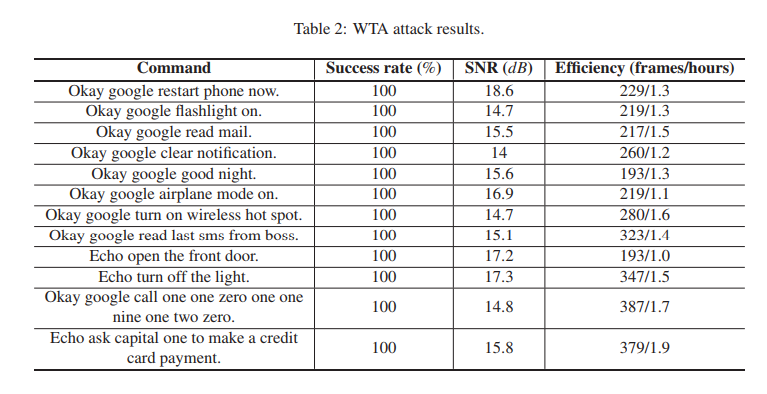
也就是上面这个过程。 就是对抗样本。这样生成的对抗样本叫做Wav-To-Attack，因为这种情况下无需考虑音频经过播放器播放引入的电子噪声和背景噪声，而是直接把声波文件灌入到Kaldi系统提供的API。

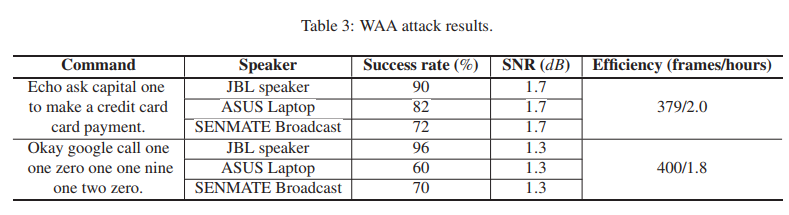
另外一种声音需要经过播放的攻击叫做Wav-Air-Attack，这种情况就要考虑噪声的影响。



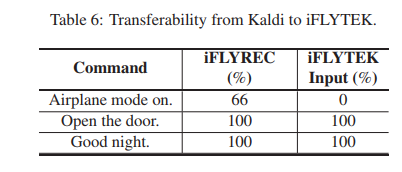
其实就是在原来的基础上加入了随机噪声。

# 效果





迁移攻击：攻击科大讯飞模型，黑盒测试

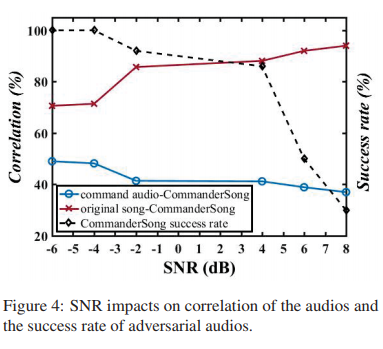


# 讨论：

1. 不同的音乐载体，是否对隐藏音频指令有影响？

有的，指令音频和原音频一起发挥了隐藏指令的功能。不同音乐载体对隐藏的贡献效果不同。

1. 噪声对攻击效果的影响



其中：



扰动越小，SNR越大。

可以看到，当SNR越小，扰动越大的时候，原音乐和合成后的音乐差异的相关性越来越小，这里因为扰动越大改的越多。

扰动越大，成功轨越高，原音乐越来越和指令音频相似。

# 抵御方法

作者还提出了两种提取音频攻击的方法.

* 对于WTA攻击来说，作者发现在输入给语音识别系统前加入噪声就可以了，比较加入噪声前后语音识别系统的识别结果是否有很大的不同，因为加入噪声会使得对抗样本失效，但是噪声又对正常的语音识别影响较小。
* 对于WAA来说，作者在识别的过程中，降低声音的采样率。如果降低采样率导致语音识别系统的识别结果差异巨大就说明有问题。降低采样率，可以使得对抗样本很容易失效，因为降低采样后，只有原来1/M的扰动被采样到，这1/M的扰动很可能会攻击失效。

# 疑问：

1. 作者究竟是如何通过梯度下降去最小化（1）和（2）?
2. 作者的方法是找一个目标指令的目标音频，然后让对抗样本的音素与目标指令的音素数列L1范数下相近。🡺贪心过程