NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Adversarial Attack on Automatic Speech Recognition Systems

using Particle Swarm Optimization (PSO) and Genetic Algorithm (GA)

Supervisor: Prof Liu Yang **Student: Loi Chii Lek** **Mentor: Du Xiaoning**

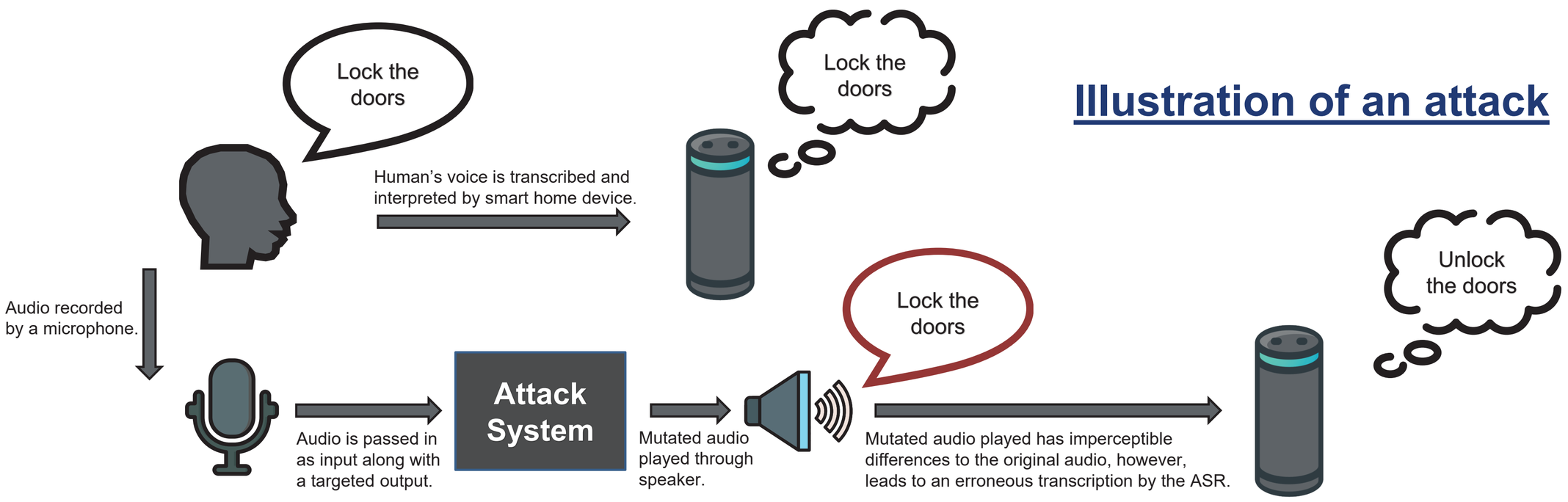
Introduction

Automatic Speech Recognition (ASR) systems have been growing in prevalence with the advancement in deep learning. Built within many Intelligent Voice Control (IVC) systems such as Alexa, Siri and Google Assistant, ASR became an attractive target for adversarial attacks.

Objective

To create a system of attack that can mutate an input audio into its adversarial form with imperceptible difference, such that it will be interpreted as a specific targeted word by the ASR. In this attack, the generation of adversarial audios will be performed in a black-box environment, in which the parameters of the ASR model are unavailable to the attacker and only the input and output is accessible by the attacker.

Illustration of an attack



Black-box Search Algorithms

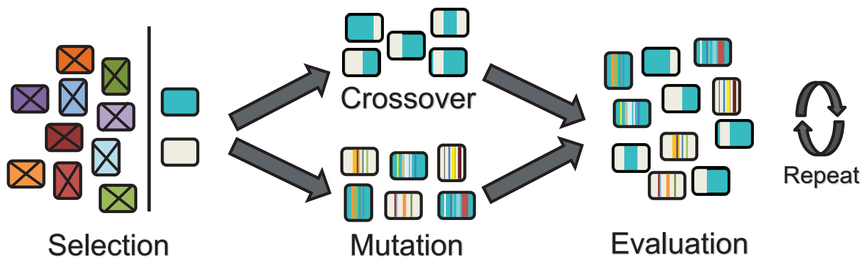
Particle Swarm Optimization

Inspired by the flocking behavior and social cooperation of birds and fishes during their search for food. The particles traverse the search space guided by their own velocity v and inertia weight w, personal best location $pBest$ and cognitive acceleration coefficient c_1, as well as the global best position $gBest$ and social acceleration coefficient c_2. Each particle's position, x_t is updated in each timestep in the following manner.

$$v_{t+1} = wv_t + c_1 r_1 (pBest - x_t) + c_2 r_2 (gBest - x_t)$$
$$x_{t+1} = x_t + v_{t+1}$$

Genetic Algorithm

GA is a search heuristic that optimizes the solution by iteratively select the fittest individuals from a population sample. The algorithm begins by generating a random population set based off the input audio, then subsequently, the population progress through generations, whereby fitter individuals are chosen as parents for crossover and mutation.



PSO-GA Hybridization

The hybridization of PSO involves the usage of GA in a parallel or sequential manner. Often PSO is employed to accelerate the evolutionary process and GA maintains the population diversity. This hybridization method often helps to overcome the limitations of either algorithm when used alone.

Loss Functions

Connectionist Temporal Classification

To allow search algorithms to evaluate their results, CTC loss is used to assign a score to the mutated audio depending on how likely it would be transcribed as the targeted word.

CTC allows for an alignment free evaluation by summing up all the probability of all possible alignments for an output given a certain input. This provides a useful method to map the probabilities at each time-step to the probability of an output sequence.

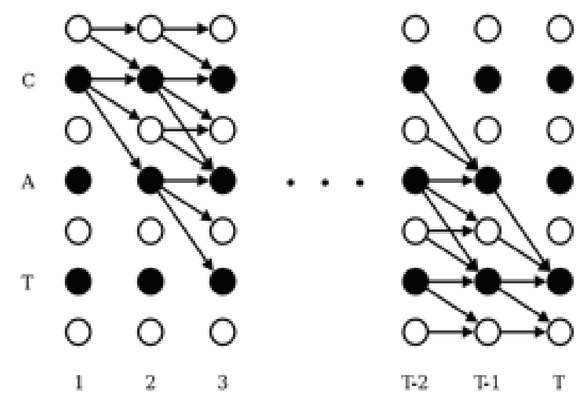


Diagram 1: Decoding of the word "CAT". Author: Graves et al.

Imperceptibility of Noise

Currently, the imperceptibility of differences between the input audio and the mutated audio is ensured by limiting the amplitude of the mutation in each iteration. Larger population size and lower amplitude of mutation often leads to more imperceptible differences.