Detecting Audio Attacks on ASR Systems with Dropout Uncertainty

Paper: https://arxiv.org/pdf/2006.01906v1.pdf

Реализовав атаку на модель детекции ключевого слова, которая используется голосовым ассистентом Alexa от Amazon, я заинтересовался темой детекции аудио атак и созданием устойчивых к атакам моделей.
Принцип аудио атаки заключается в генерации искажения с помощью дифференцируемых(с ограничениями) параметров `\theta`.

В данном исследовании реализовывалась детекция аудио атаки на end-to-end автоматических моделей распознавания аудио и перевода аудио в текст.
Идея заключается в изменении параметра dropout rate при генерации искажения, а также при детекции и наблюдении за мерой неопределенности значений спектрограммы оригинальной аудиодорожки и испорченной аудиодорожки.

Атаки удалось детектировать при разном звуковом фоне (городской шум, например).

Слабые стороны: метод детекции атаки с помощью дропаута не гарантирует точность. Он может быть как эффективен с определенной вероятностью, так и бесполезен .