Cocaine Noodles: Exploiting the Gap between human and machine speech recognition

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/b53564a1-b248-4282-9458-56eee94145d9/exploiting_the_gap_between_human_and_machine_sppech_recognition.pdf

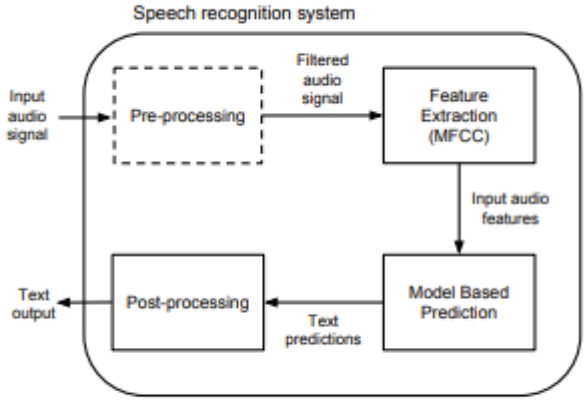

Figure 1: Workflow of a typical speech recognition system.

- 음성 인식 방식

1. pre-processing : 배경 소음 제거, 음성 인식과 관련 없는 주파수 필터링, 입력 신호 일부 제어
2. MFCC (Mel-frequency cepstral coefficients) : 음성 인식 시스템으로 주로 사용
3. hidden Markov model 이나 ann 거쳐서 text로 나옴

- MFCC
  - short-term power spectrum of audio on a nonlinear mel frequency scale

- 멜 스케일에서 주파수 간격이 동일한 청각 기반
- 각 멜 전력 로그는 이산 코사인 변환됨
- 파라미터는 음성 인식 시스템이 해당 텍스트를 정확하게 예측할 수 있도록 충분한 오디오 기능을 보존하는 방식으로 조정되지만 사람이 인식하고 이해하는 오디오 신호를 변경

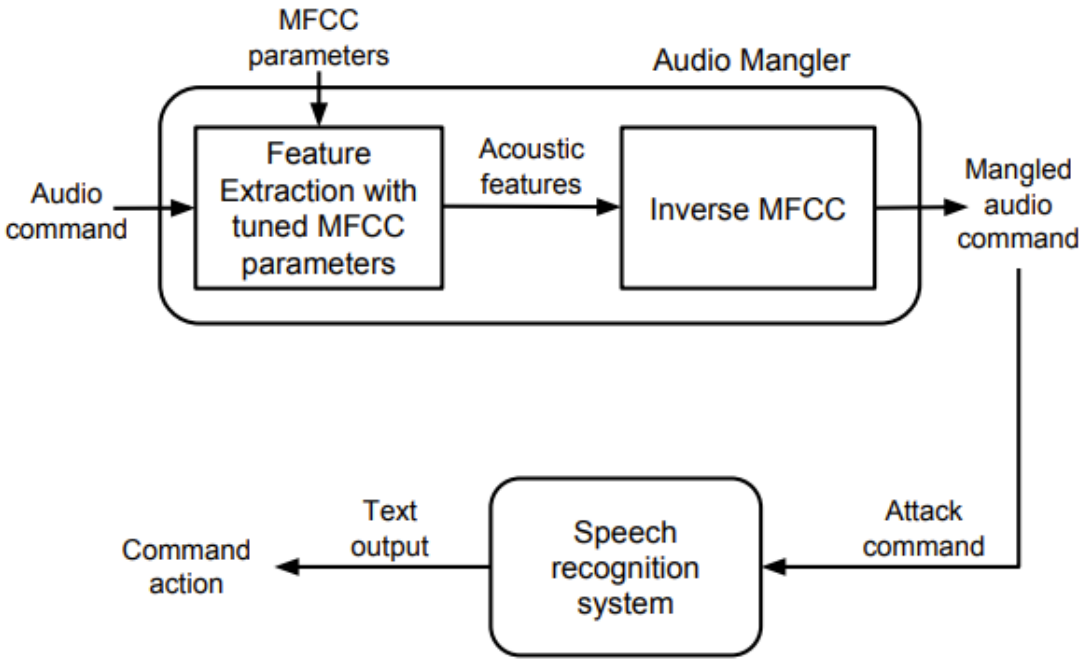

Figure 2: Attack outline.

- Audio Mangling
  - Mangle : 망치다
  - 목표 : 사람이 음성 신호를 이해하지 못하도록 망치지만, 음성 인식 시스템이 정확하게 예측할 수 있도록 하게끔
    - MFCC Parameterts
      - 4개의 독립적인 파라미터 (wintime, hoptime, numcep, nbands)
      - mangled 음성을 시스템에서 인식한다면 그 결과를 api 에 제출해서 api는 다섯가지의 가능한 경우를 보여준다. 만일 가능한 경우가 없다면 에러를

반환하는 식으로 해서 파라미터를 찾아냄

- feature extraction with tuned mfcc paramters
  - MFCC는 음성 신호 모든 것을 담지 않음
  - 공격 명령을 생성하는 데 사용되는 조정된 MFCC 매개 변수는 인간의 이해를 저해하는 방식으로 이러한 정보 손실을 더욱 증가
  - MFCC 파라미터 튜닝을 통해 mangled 음성이 타겟 음성 인식 시스템으로 정확하게 가게끔 선택함
- inverse MFCC
  - MFCC의 계산은 손싱성이 있고, 인접 지역 에너지를 통해 계산이 되기에 Inversion을 통해 원 음성 신호를 망쳐 인간이 이해할 수 없도록 만듦

- 접근 방식 : 음성 인식 시스템에 사용되는 기능이 대부분 방해를 받지 않고(추출된 후 재구성되기 때문에), 비추출된 기능은 재구성에서 손실된다는 것
- mangled 음성의 매개 변수를 조정하고 수동으로 테스트하여 목표 음성 인식 시스템의 복사본에 의해 망친 출력이 허용되는지 확인하고 다음을 확인하는 시행착오 방식으로 절차