Securitatea sistemelor de vorbire

Tudor Bumbu

Vulnerabilitatea sistemelor vocale

Unde apar riscurile majore?

- Transcrieri intenționat greșite sau comenzi ascunse.
- Executarea unor acțiuni nedorite
  (de ex. *comenzi smart home nesolicitate*).
- În biometria vocală sistemul poate accepta o voce falsificată în locul utilizatorului real.

Trei clase majore de amenințări

Riscurile de securitate se împart în trei mari categorii:

1. Atacuri adversariale în sistemele ASR
Un semnal audio modificat subtil. Pentru om sună perfect normal, dar forțează sistemul ASR să transcrie complet altceva.

2. Replay
Atacatorul folosește o înregistrare reală a vocii victimei și o redă sistemului dintr-un difuzor.

3. Spoofing
Generarea artificială a unei voci (clonare) folosind tehnici de sinteză a vorbirii (TTS).

Apărarea de atacuri

Securitatea eficientă se bazează pe ***logică***.

Apărarea contra Replay

- Verifică dacă vocea aparține unei persoane reale prezente fizic. Metoda *Challenge-Response* cere utilizatorului să citească o frază generată aleator pe loc, făcând inutile înregistrările vechi.

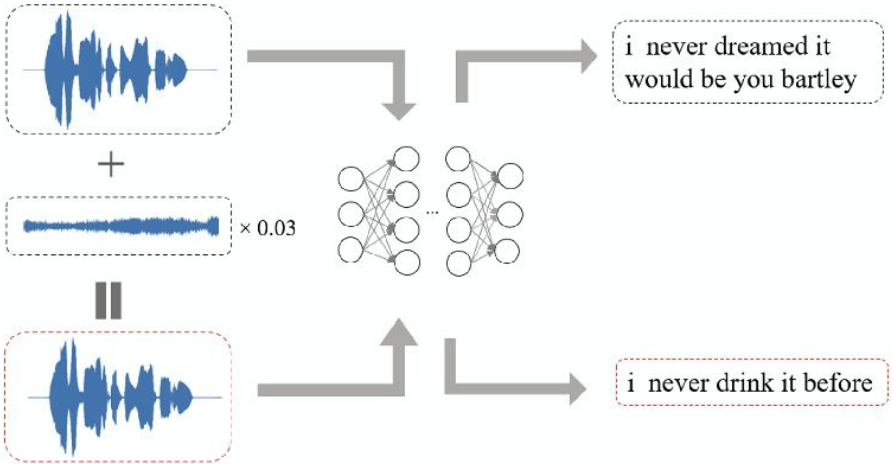
Modele Anti-Spoofing contra Deepfake audio

- Algoritmi bazați pe rețele neuronale care analizează artefactele acustice invizibile pentru a face diferența între o voce biologică reală și una sintetizată.

Antrenare adversarială

Cum protejăm modelul ASR propriu-zis de atacurile care încearcă să îl păcălească?

- În timpul antrenării, modelului i se expun intenționat exemple audio perturbate (atacate). Astfel, rețeaua învață să fie robustă și să ignore zgomotele malițioase.
- Preprocesarea audio înainte de a ajunge în model, pentru a curăța frecvențele suspecte, având grijă să nu degradăm calitatea vorbirii normale.
- Monitorizarea comportamentului și detectarea intrărilor atipice.



Măsuri operaționale pentru biometria vocală

Vocea singură nu trebuie să fie unicul factor de securitate pentru acțiuni critice.

Măsuri:

- MFA (Autentificare Multi-Factor).
- Sistemul vocal execută acțiunea doar după o confirmare manuală pe ecran.
- Limitarea numărului de încercări eșuate pentru a preveni atacurile de forță brută.
- Monitorizarea activităților pentru detectarea fraudelor.