# Atacuri asupra Sistemelor de Vorbire

## 12.1 Introducere: Noul Peisaj al Amenințărilor Vocale

Odată cu integrarea profundă a sistemelor vocale în aplicații comerciale și critice, de la asistenți digitali care controlează case inteligente la autentificarea biometrică pentru tranzacții bancare, securitatea acestor sisteme a devenit o preocupare majoră. Interfața vocală, prin natura sa deschisă și accesibilă, introduce vectori de atac unici. Integritatea, disponibilitatea și confidențialitatea sistemelor automate de recunoaștere a vorbirii (ASR) și de verificare a vorbitorului (ASV) sunt acum ținte directe pentru atacuri malițioase. Acest capitol abordează două dintre cele mai semnificative și studiate categorii de amenințări: **atacurile adversariale** (adversarial attacks), care vizează în principal ASR, și **atacurile de tip replay și spoofing**, care țintesc sistemele de verificare a vorbitorului .

## 12.2 Atacuri Adversariale asupra Recunoașterii Vorbirii

### 12.2.1. Definiție și Mecanism de Atac

Atacurile adversariale, în contextul audio, implică adăugarea unor perturbații intenționate și atent optimizate la un semnal vocal. Aceste perturbații sunt adesea calculate pentru a fi cvasi-imperceptibile pentru sistemul auditiv uman, dar sunt concepute special pentru a exploata vulnerabilitățile modelelor de deep learning care stau la baza sistemelor ASR moderne. Scopul nu este simpla degradare a semnalului, ci inducerea în eroare a modelului într-un mod controlat.

Lucrarea fundamentală a lui Carlini și Wagner (2018) a demonstrat nu doar posibilitatea, ci și eficacitatea acestor atacuri. Ei au arătat cum un semnal audio benign poate fi modificat astfel încât, deși pentru un om sună normal, un sistem ASR de ultimă generație, cum ar fi DeepSpeech, transcrie o frază țintă complet diferită, aleasă de atacator. Aceasta transformă atacul dintr-o simplă eroare de recunoaștere într-o injecție de comandă ascunsă.

### 12.2.2. Vectori de Atac și Scenarii de Risc

Atacurile adversariale pot fi clasificate în funcție de modul de livrare. Atacurile "white-box" presupun că atacatorul are acces deplin la arhitectura și ponderile modelului, permițând optimizarea precisă a perturbației. Mai îngrijorătoare sunt atacurile "black-box", unde atacatorul poate interoga modelul doar prin API și estimează gradientul.

Scenariile de risc devin concrete atunci când aceste atacuri sunt transferate din domeniul digital în cel fizic ("over-the-air"). Cercetările au demonstrat că perturbațiile adversariale pot fi redate prin difuzoare într-o cameră și pot păcăli cu succes un microfon aflat la distanță (de exemplu, un asistent vocal). Hussain et al. (2021) au explorat înțelegerea și atenuarea acestor exemple adversariale audio, subliniind robustețea lor în scenarii reale. Atacurile pot fi rafinate ulterior folosind modele psihoacustice pentru a ascunde perturbațiile în frecvențe unde auzul uman este mai puțin sensibil.

Un exemplu clar de risc este comanda unui asistent vocal pentru a efectua acțiuni cu impact asupra securității ("deschide ușa", "dezactivează alarma", "transferă bani"), în timp ce utilizatorul legitim percepe doar o conversație benignă sau zgomot de fond.

### 12.2.3. Măsuri de Apărare și Limitări

Apărarea împotriva atacurilor adversariale este o problemă deschisă în cercetare. Soluțiile propuse includ **antrenarea adversarială (adversarial training)**, în care modelul este re-antrenat pe un set de date care include exemple de semnale perturbate, forțându-l să învețe invarianța la aceste perturbații. Alte tehnici implică **pre-procesarea semnalului**, cum ar fi filtrarea benzilor de frecvență sau adăugarea de zgomot aleatoriu pentru a distruge structura perturbației. Detectarea anomaliilor în straturile de ieșire ale modelului (de exemplu, "logit noising") este o altă abordare. Totuși, majoritatea acestor apărări implică un compromis: ele pot reduce eficacitatea atacurilor, dar adesea degradează și performanța generală a sistemului ASR pe semnale legitime, sau pot fi depășite de atacuri adaptive care sunt concepute special pentru a ocoli apărarea respectivă.

## 12.3 Atacuri de Tip Replay și Spoofing asupra Verificării Vorbitorului

Dacă atacurile adversariale vizează ASR, atacurile de tip replay și spoofing vizează în mod specific sistemele de **autentificare biometrică vocală (ASV)**.

### 12.3.1 Clasificarea Amenințărilor

Aceste atacuri, cunoscute sub denumirea de "presentation attacks", sunt clasificate în funcție de artefactul prezentat senzorului (microfonului):

1. **Atacuri de Tip Replay (Redare):** Aceasta este cea mai simplă formă de atac, în care atacatorul redă o înregistrare reală a vocii utilizatorului legitim. Înregistrarea poate fi obținută prin interceptare, din mesaje vocale sau chiar din conținut public (de exemplu, rețele sociale, conferințe).
2. **Atacuri de Tip Spoofing (Falsificare):** Acestea sunt mult mai sofisticate și implică generarea unui semnal audio sintetic care imită vocea țintei. Ele se împart în două categorii principale:
   - **Sinteza Vocală (Text-to-Speech, TTS):** Utilizarea unui model TTS antrenat pe vocea țintei pentru a genera orice frază dorită.
   - **Conversia Vocală (Voice Conversion, VC):** Utilizarea unui model care transformă vocea atacatorului pentru a suna ca cea a țintei.

Ambele forme de spoofing sunt adesea denumite **deepfakes audio** și devin din ce în ce mai accesibile și realiste. Pericolul constă în faptul că semnalul falsificat conține caracteristicile biometrice esențiale pe care sistemul ASV este antrenat să le recunoască, făcând detecția extrem de dificilă. Aceste atacuri sunt deosebit de periculoase în aplicații critice precum accesul la conturi bancare sau controlul vocal al dispozitivelor IoT.

### 12.3.2 Contramăsuri (CM) și Detecția Falsificărilor

Răspunsul tehnic la amenințările de tip replay și spoofing constă în dezvoltarea de **Contramăsuri (Countermeasures - CM)**. În esență, un sistem CM este un clasificator binar antrenat să facă distincția între vorbirea autentică, numită **"bona fide"**, și un semnal de atac, numit **"spoofed"**. Provocarea fundamentală constă în faptul că un atac de succes este, prin definiție, proiectat să semene biometric cu ținta, făcând ca sistemul ASV (verificarea vorbitorului) să fie inutil pentru propria sa

apărare. O analiză realizată de Nandwana et al. (2023) a arătat că multe sisteme ASV pot fi înșelate prin metode simple de redare audio. Prin urmare, contramăsurile trebuie să caute dovezi *ortogonale* față de identitatea biometrică a vorbitorului. Aceste dovezi se încadrează, în general, în două categorii: analiza artefactelor acustice și detecția de viață.

**Analiza Artefactelor Acustice**

Această abordare se bazează pe premisa că procesul de redare sau sinteză, indiferent cât de avansat, lasă "urme" sau artefacte fizice în semnalul audio, pe care un tract vocal uman nu le produce.

**Pentru Atacurile de Tip Replay**: Artefactele sunt introduse de lanțul electro-acustic. Semnalul audio redat este capturat de un microfon, iar acest proces introduce distorsiuni neliniare, răspunsul în frecvență specific al difuzorului, reverberațiile camerei și zgomotul de fond specific microfonului de înregistrare. Un sistem CM poate fi antrenat să detecteze aceste distorsiuni de canal subtile, care sunt absente într-o înregistrare "live" de înaltă calitate.

**Pentru Atacurile de Tip Spoofing (TTS/VC):** Artefactele sunt de natură algoritmică, introduse de procesul de generare a vorbirii. Modelele timpurii de sinteză parametrică sufereau de o "netezime" spectrală excesivă (oversmoothing), creând o voce robotică, ușor de detectat. Deși "deepfakes" moderne bazate pe GAN-uri sunt mult mai realiste, ele încă pot introduce imperfecțiuni, cum ar fi discontinuități de fază la joncțiunea cadrelor audio sau un zgomot de fond algoritmic specific vocoderului.

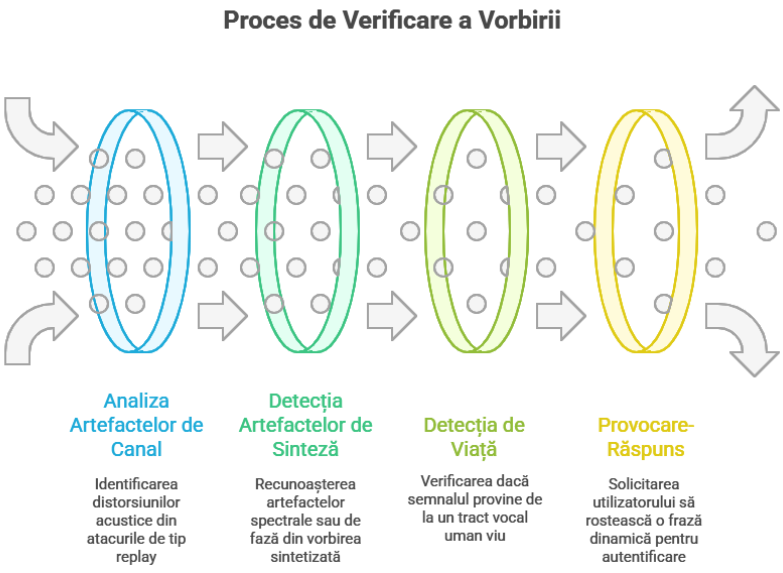


**Fig. 4. Procesul de verificare a vorbirii**

Pentru a captura aceste artefacte extrem de fine (fig.4) , trăsăturile acustice standard folosite în ASR și ASV, cum ar fi MFCC (Coeficienții Cepstrali pe Scara Mel), s-au dovedit a fi suboptime. Cercetarea a demonstrat că trăsăturile care operează pe o scară liniară sau cu o rezoluție diferită sunt mult mai eficiente. De exemplu, **Coeficienții Cepstrali pe Frecvență Liniară (LFCC)** sau **Coeficienții Cepstrali pe Scara Constantă Q (CQCC)** s-au dovedit a fi semnificativ mai buni la

discriminarea între semnalele autentice și cele sintetice, deoarece anomaliile de sinteză sunt adesea mai evidente în afara benzilor de frecvență Mel (Sahidullah et al., 2015).

### Detecția de Viață (Liveness Detection)

Spre deosebire de analiza artefactelor, detecția de viață nu caută semne de *mașină,* ci semne de *viață*—dovezi fiziologice ale unui tract vocal uman. Acestea includ evenimente acustice non-lingvistice pe care un om le produce în mod natural și inconștient în timpul vorbirii, dar pe care un sintetizator (care este antrenat de obicei doar pe text curat) le-ar omite. Exemple de astfel de semnale includ zgomotul specific al respirației între cuvinte, clicurile salivare, sau micro-variațiile ascuțite (pop-uri) asociate sunetelor plozive (de exemplu, 'p', 't', 'k') (Kinnunen et al., 2017). Detectarea prezenței acestor indicatori fiziologici poate oferi o dovadă puternică că semnalul provine de la un vorbitor uman autentic.

### Metode Active (Challenge-Response)

O contramăsură activă, spre deosebire de cele pasive de mai sus, este mecanismul de **provocare-răspuns (challenge-response)**. În acest scenariu, sistemul de autentificare nu solicită o parolă statică (de exemplu, "Parola mea vocală este..."), ci generează o frază unică, aleatorie, la momentul autentificării (de exemplu, "Pisica verde a sărit peste lună") și solicită utilizatorului să o rostească. Această metodă este extrem de eficientă pentru a învinge *toate* atacurile de tip replay, deoarece este practic imposibil ca un atacator să fi pre-înregistrat acea secvență specifică. Totuși, această metodă rămâne vulnerabilă la un atac de spoofing sofisticat în timp real, în care un sistem TTS/VC interceptează provocarea și generează răspunsul falsificat aproape instantaneu.

### Standardizare și Benchmarking: Provocarea ASVspoof

Progresul rapid în domeniul contramăsurilor a fost catalizat și standardizat de seria de provocări academice **ASVspoof** (Wu et al., 2015; Kinnunen et al., 2017). Începând cu 2015, această inițiativă a furnizat comunității științifice seturi de date la scară largă, partajate public, care conțin atât vorbire *bona fide*, cât și o gamă extrem de variată și evolutivă de atacuri de spoofing (diferite tehnici de sinteză, conversie vocală și redare în diverse condiții).

Prin stabilirea unui protocol de testare comun și a unor metrici standardizate (în special **Equal Error Rate - EER** pentru contramăsuri), ASVspoof a permis cercetătorilor din întreaga lume să compare în mod fiabil eficacitatea noilor algoritmi de detecție. Această standardizare a fost crucială pentru a identifica rapid ce tehnici funcționează (de exemplu, trecerea de la MFCC la LFCC/CQCC) și pentru a expune vulnerabilitățile la noi tipuri de atacuri (de exemplu, atacuri deepfake pe care modelele mai vechi nu le puteau detecta).

## 12.4 Concluzii

Securitatea sistemelor vocale este o provocare multidimensională care evoluează rapid. Atacurile, atât cele adversariale împotriva ASR, cât și cele de spoofing împotriva ASV, devin din ce în ce mai sofisticate. Aceasta impune o schimbare de paradigmă în proiectarea sistemelor: modelele

trebuie optimizate nu doar pentru acuratețe, ci și pentru **reziliență la manipulare**. Sunt necesare metode de apărare mai robuste, standarde comune de evaluare (precum cele promovate de ASVspoof) și o abordare proactivă, "security-by-design", în întregul ciclu de viață al dezvoltării sistemelor vocale.