Voice Commands for Robot Orientation in Space

Oneata Razvan-Mihai, Troaca Viorel Mario and **Ghimpeteanu Alexandru Andrei**
Universitatea din București
razvan-mihai.oneata@s.unibuc.ro
viorel-mario.troaca@s.unibuc.ro
alexandru-andrei.ghimpeteanu@s.unibuc.ro

Abstract

Acest document prezintă dezvoltarea unui sistem de control vocal pentru orientarea roboților mobili, eliminând necesitatea telecomenzilor fizice. Utilizând platforma Edge Impulse, implementăm o soluție de tip TinyML pentru a detecta comenzi vocale (Keyword Spotting) direct pe hardware-ul robotului, asigurând o latență minimă. Noutatea propunerii constă în integrarea clasificării audio cu un sistem de navigație inerțială (IMU) și un controler PID, permițând corecția activă a orientării în spațiu. Raportul descrie setul de date utilizat, arhitectura rețelei neurale convoluționale (CNN) și planul de evaluare a performanței sistemului.

1 Introduction

Interacțiunea om-robot (Human-Robot Interaction - HRI) a devenit un domeniu fundamental în robotică, având ca scop dezvoltarea unor interfețe naturale și intuitive. În timp ce metodele tradiționale de control implică utilizarea unor dispozitive fizice, precum telecomenzi sau joystickuri, acestea limitează libertatea de mișcare a operatorului.

În acest proiect, propunem un sistem intitulat "Voice Commands for Robot Orientation in Space", care elimină necesitatea perifericelor fizice. Obiectivul principal este dezvoltarea unui robot mobil capabil să își modifice orientarea precis folosind comenzi vocale în limbaj natural.

Pentru a atinge acest scop într-un mod eficient, utilizăm platforma **Edge Impulse**, care ne permite să implementăm algoritmi de învățare automată (Machine Learning) direct pe hardware-ul robotului (TinyML). Această abordare elimină latența introdusă de procesarea în cloud și asigură un răspuns rapid la comenzi precum "Left", "Right" sau "Stop".

Noutatea sistemului constă în integrarea modelului audio dezvoltat în Edge Impulse cu un sistem de navigație inerțială. Folosind datele de la un senzor IMU, robotul nu execută doar o simplă comandă motorie, ci își corectează orientarea în timp real, oprindu-se exact la unghiul dorit, independent de alunecarea roților sau de suprafața de rulare.

2 Related Work

O evoluție majoră în domeniul inteligenței artificiale este apariția conceptului de *TinyML*, care permite rularea modelelor de învățare automată pe microcontrollere cu resurse limitate. Această abordare este critică pentru roboții mobili, unde dependența de o conexiune la internet pentru procesare în cloud ar introduce latență inacceptabilă. Platforma **Edge Impulse** facilitează această tranziție, oferind un flux de lucru optimizat pentru implementarea rețelelor neurale pe dispozitive embedded.

În ceea ce privește arhitectura modelelor audio, Sainath and Parada (2015) au demonstrat că Rețelele Neurale Convoluționale (CNN) sunt mult mai eficiente decât modelele tradiționale pentru sarcina de *Keyword Spotting* (KWS). Arhitecturile moderne utilizate de Edge Impulse se bazează pe aceste principii, analizând spectrogramele semnalelor audio pentru a extrage trăsături relevante cu un consum minim de energie.

Pentru antrenarea și validarea modelelor, Warden (2018) a introdus setul de date *Google Speech Commands*. Acesta a devenit standardul de referință în domeniu, oferind mii de exemple pentru comenzi scurte (precum "Left", "Right", "Go"), fiind baza pe care construim și setul nostru de date.

Deși există implementări de control vocal, majoritatea funcționează în buclă deschisă (*open-loop*), unde robotul execută comanda fără a valida rezultatul mișcării. Proiectul nostru propune o îmbunătățire prin integrarea unui senzor inerțial (IMU) în bucla de control, asigurând o corespondență precisă între comanda vocală și orientarea fizică finală.

3 Methodology

Metodologia propusă integrează procesarea audio *on-device* cu controlul în buclă închisă al mișcării robotului. Procesul este divizat în trei etape: pregătirea datelor, antrenarea modelului neuronal și controlul navigației.

3.1 Dataset and Exploratory Analysis

Pentru antrenarea modelului, utilizăm setul de date *Google Speech Commands*, selectând patru clase principale: "Left", "Right", "Go" și "Stop", la care se adaugă clasele auxiliare "Unknown" și "Noise".

Pentru a asigura robustețea în mediul real, am extins acest set cu date proprii (*custom dataset*), înregistrate folosind microfonul robotului. Am efectuat o analiză exploratorie a datelor (EDA) utilizând instrumentele de vizualizare din Edge Impulse. Aceasta ne-a permis să identificăm și să eliminăm eșantioanele cu raport semnal-zgomot (SNR) scăzut și să echilibrăm numărul de înregistrări per clasă, prevenind astfel bias-ul modelului către o anumită comandă.

3.2 Model Architecture

Procesarea semnalului începe cu un bloc DSP (*Digital Signal Processing*) care extrage spectrograme MFE (*Mel-filterbank Energy*). Acestea servesc drept date de intrare pentru o Rețea Neurală Convoluțională (CNN) 2D.

Alegerea unui CNN este motivată de capacitatea acestuia de a detecta tipare invariante în timp și frecvență. Modelul este optimizat prin tehnici de *quantization* (int8), reducând dimensiunea și timpul de inferență pentru a permite rularea pe microcontroller.

3.3 Robot Control Strategy

Odată clasificată o comandă vocală cu o probabilitate ce depășește un prag de încredere (ex. 0.8), sistemul activează controlerul de mișcare.

Spre deosebire de o abordare bazată pe timp (ex. "rotește motorul 1 secundă"), implementăm un algoritm PID (*Proportional-Integral-Derivative*) care citește datele de la giroscopul senzorului IMU. Pentru comenzile "Left" și "Right", PID-ul monitorizează unghiul de girație (*yaw*) și ajustează puterea motoarelor în timp real până când eroarea dintre orientarea curentă și cea țintă (90 grade) devine zero.

4 Evaluation Plan

Pentru validarea sistemului propus, am definit două seturi de metrici de performanță: unul pentru clasificatorul audio și unul pentru execuția fizică a comenzilor.

4.1 Audio Model Metrics

Performanța modelului antrenat în Edge Impulse va fi evaluată pe un set de testare separat (20% din date), pe care rețeaua nu l-a văzut în timpul antrenării. Vom urmări:

- **Acuratețea globală:** Procentul de comenzi clasificate corect.
- **Matricea de Confuzie:** Esențială pentru a identifica erorile critice (ex. confundarea comenzii "Left" cu "Right").
- **Performanța la zgomot:** Vom compara acuratețea modelului în două scenarii: mediu silențios vs. mediu cu zgomot ambiental, pentru a testa robustețea algoritmului.

4.2 Robot Navigation Accuracy

Evaluarea fizică va măsura capacitatea sistemului de control (IMU + PID) de a executa comenzile primite. Vom efectua câte 10 teste pentru fiecare comandă de orientare (ex. rotire 90 de grade) și vom măsura:

- **Eroarea absolută medie:** Diferența dintre unghiul țintă și orientarea finală a robotului măsurată de senzorii externi.
- **Timpul de răspuns (Latența):** Durata de timp scursă între momentul finalizării comenzii vocale și inițierea mișcării motoarelor. Ne propunem o latență sub 500ms pentru o experiență de utilizare fluidă.

5 Conclusion

În această etapă a proiectului, am definit arhitectura sistemului, am selectat setul de date (Google Speech Commands) și am stabilit fluxul de lucru folosind platforma Edge Impulse pentru o implementare TinyML eficientă. Următorii pași includ antrenarea finală a modelului CNN, exportarea bibliotecii C++ pe microcontroller și calibrarea buclei de control PID pentru a asigura o navigație precisă în spațiu.

References

Tara N Sainath and Carolina Parada. 2015. Convolutional neural networks for small-footprint keyword spotting. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.