CIDAI Centre of Innovation for Data tech and Artificial Intelligence

Informe de prova de concepte CIDAI-POC-2021-05

Assistent de veu “light”

Sistema autònom

Informe elaborat per:

Sergi Mercadé Laborda
Sergi Sánchez Deutsch
Josep Escrig
Àngel Martín

i2cat

CIDAI Centre of Innovation for Data tech and Artificial Intelligence

i2cat

CIDAI-POC-2021-05

Cita suggerida. CIDAI-POC-2021-05 // Assistent de veu "light" Sistema autònom per Sergi Mercadé Laborda, Sergi Sánchez Deutsch, Josep Escrig, Àngel Martín, CIDAI, 2021. Llicencia: CC BY-NC-SA 4.0

Índex

1. Resum

1.1. Objectius

Els assistents de veu personals s'han popularitzat enormement en els últims anys per dur a terme una gran varietat de tasques quotidianes, incloses les que involucren el comandament de dispositius intel·ligents dins d'un habitatge. Mentre la funció principal d'aquests assistents és facilitar la vida diària de les persones, la seva utilització sovint planteja un problema de privacitat derivat del processament de la veu al núvol.

L'objectiu d'aquest projecte és la creació d'un sistema senzill de reconeixement de paraules clau i comandes capaç de funcionar íntegrament en local, és a dir, sense dependre en cap moment de connexió a internet ni processament en servidors externs. El sistema està pensat per funcionar en dispositius de la llar amb recursos limitats i fa ús d'un algoritme d'intel·ligència artificial que pot ser entrenat per a detectar paraules clau personalitzades.

1.2. Visió

La intel·ligència ambiental fa referència a aquells entorns que responen a la presència i les accions dels humans mitjançant dispositius electrònics. Els assistents de veu són un clar exemple d'aquest tipus d'intel·ligència que s'han popularitzat enormement. Són sistemes capaços de reconèixer un context i adaptar-se com a resposta a les comandes de veu, a més de ser personalitzables d'acord a les nostres necessitats. Acostumen a funcionar fent servir la computació al núvol, mitjançant la qual el processament de les entrades es realitza en servidors externs.

1.3. Resum del problema

Els assistents de veu personals són àmpliament coneguts i utilitzats en el dia a dia. El seu funcionament acostuma a tenir una dependència total de la connexió a internet ja que el processament de veu es fa en servidors del fabricant del dispositiu. Aquest fet genera sovint dubtes sobre la privacitat i la seguretat de les dades que s'envien, a més d'incrementar el temps de resposta i el consum energètic total de l'aplicació. Aquesta connexió, però, es pot evitar per dur a terme algunes comandes senzilles que no requereixin d'una gran capacitat de processament en un servidor extern.

1.4. Solució

Es proposa un sistema basat en intel·ligència artificial que permet incorporar el reconeixement de veu en dispositius de la llar com una nevera, una rentadora o una torradora que tenen un interfície d'usuari limitat (llums i botons), que fan servir dispositius controladors de baix cost i amb una capacitat de processament limitada. El sistema ha de permetre detectar una paraula clau seguida d'una comanda senzilla fent tot el processament necessari en local.

2. Descripció del problema que resol

Actualment molts habitatges incorporen dispositius intel·ligents com aspiradores, cuines o llums. Aquests dispositius acostumen a estar integrats amb assistents de veu personals, fet que permet controlar-los fent servir comandes de veu. Per tal que aquesta integració funcioni és necessària una connexió a internet, ja que el processament de les dades de veu se sol dur a terme en un servidor extern.

Tot i que l'anonimització de dades és una pràctica habitual en aquest tipus de sistemes, molts usuaris encara tenen dubtes sobre com es tracten els temes de la privacitat i la seguretat d'aquestes dades de veu. Aquesta dependència amb el cloud també fa que l'usuari estigui subjecte a uns termes i condicions del fabricant respecte les seves dades que sovint són canviants i poc clares.

Per a l'accionament d'alguns dispositius intel·ligents les comandes de veu solen ser senzilles. En aquests casos, actualment es pot disposar de processadors de baix cost que permeten fer tot el tractament de la veu necessari de forma local sense necessitat d'una connexió a internet. Això permet una millora respecte la privacitat de l'usuari, a més de fer el sistema més robust davant atacs externs que podrien arribar per internet.

3. Implementació de la solució

3.1. Arquitectura, tecnologia i dades utilitzades

Arquitectura

El sistema està format per dos tipus de dispositius diferents:

- Dispositius controladors de baix cost. En la prova de concepte s'han fet servir dispositius amb micròfon integrat i connexió BLE. La funció d'aquests dispositius és captar el soroll ambiental, realitzar un processament inicial per detectar si l'àudio captat es correspon a soroll o veu i controlar altres dispositius no intel·ligents connectats.
- Un dispositiu central. En la prova de concepte s'ha fet servir una Raspberry Pi 3B. Aquest dispositiu central s'encarrega de gestionar tot el sistema. Controla els dispositius de baix cost i s'encarrega de dur a terme el processament de la detecció de paraules clau i comandes en local, sense dependre de cap servei al núvol.

El sistema no està lligat a una plataforma de hardware en concret. Tots els dispositius de baix cost estan connectats al dispositiu central mitjançant BLE. Els dispositius de baix cost estan associats a dispositius que tinguem a la llar. Es probable que la comanda de veu la rebi mes d'un dispositiu. Quan els dispositius d'una zona detecten un soroll, realitzen un primer processat en local per decidir si es tracta d'un soroll ambiental, un esdeveniment a detectar o veu. En cas de ser un esdeveniment a detectar o veu, els dispositius avisen al dispositiu central que han detectat informació rellevant, i el dispositiu central decideix quin dispositiu controlador li envia la informació i notifica al controlador corresponent. El controlador seleccionat envia tot l'àudio rellevant que ha captat i el dispositiu central realitza un processat en local per tal d'extreure paraules clau i comandes del sistema. Un cop s'ha processat l'àudio, el dispositiu central duu a terme les accions corresponents a la comanda o l'esdeveniment detectat.

Tecnologia

La detecció de paraules clau i de diferents esdeveniments es realitza en local en el dispositiu central mitjançant xarxes neuronals convolucionals. Aquest tipus de xarxes neuronals estan especialitzades en la detecció de patrons, especialment en imatges. Mitjançant el càlcul de l'espectrograma dels clips d'àudio, s'obté una "imatge" de les components freqüencials de l'àudio en el temps. Aquesta imatge és introduïda a la xarxa neuronal convolucional, i com a resultat s'obtenen les probabilitats que les diferents paraules claus o esdeveniments apareguin dins del segment d'àudio analitzat. A la figura 1 es mostra un esquema d'aquest procediment.

La xarxa neuronal convolucional utilitzada té una estructura formada principalment per:

- 4 capes convolucionals, encarregades d'analitzar l'espectograma i fer una extracció de patrons que facilitin la classificació de la imatge inicial. També s'aplica una capa d'activació "ReLU" i una amb "Max Pooling" després de cada capa convolucional.
- 4 capes "fully-connected" que redueixen la dimensionalitat de les dades i generen un vector amb la probabilitat de que l'espectograma correspongui a cadascuna de les classes. Després de cada capa "fully-connected" (excepte l'última) s'aplica una capa de normalització, una d'activació "ReLU" i una capa amb "Dropout" per tal d'evitar un problema d'overfitting durant el procés d'entrenament.

Dades

Les dades utilitzades per a entrenar la xarxa neuronal en aquest projecte són petits clips d'àudio d'un segon de durada. Gran part de les dades que s'han fet servir per entrenar el model de xarxa neuronal provenen d'una base de dades oberta publicada per Google per fer recerca, anomenada Speech Commands Dataset.

Per tal d'entrenar el sistema a reconèixer les diferents paraules clau escollides per a la prova de concepte, l'equip d'I2CAT ha creat una petita base de dades addicional amb exemples de paraules clau a detectar. El conjunt de mostres està format per veus naturals, enregistrades per dones i homes voluntaris de diferents edats, i veus sintètiques d'alta qualitat, generades mitjançant sistemes de síntesis de veu.

Per tal d'entrenar el model d'intel·ligència artificial, l'equip d'I2CAT ha combinat les dues bases de dades i ha fet servir tècniques d'augmentació de dades per a generar noves mostres a partir de les mostres existents, incrementant de manera sintètica la varietat de les dades i el nombre de mostres total.

S'han utilitzat dues tècniques principals per dur a terme l'augment de dades de forma sintètica:

- Alteració de la freqüència fonamental (pitch), fet que permet obtenir un mateix àudio amb diferents percepcions freqüencials. El resultat és un espectrograma que conté la mateixa paraula clau però amb diferents freqüències, fet que es pot veure com si fos una mostra gravada per un altre interlocutor .
- Addició de soroll de fons. A més dels àudios amb paraules claus, també s'han gravat mostres amb diferents tipus de soroll per tal d'afegir-los als arxius originals i obtenir més varietat de situacions.

Aquestes dues tècniques s'apliquen de forma aleatòria als àudios originals de paraules clau, tant individualment com de forma conjunta, per tal d'aconseguir un número de mostres que estigui balancejat amb la resta de classes del dataset.

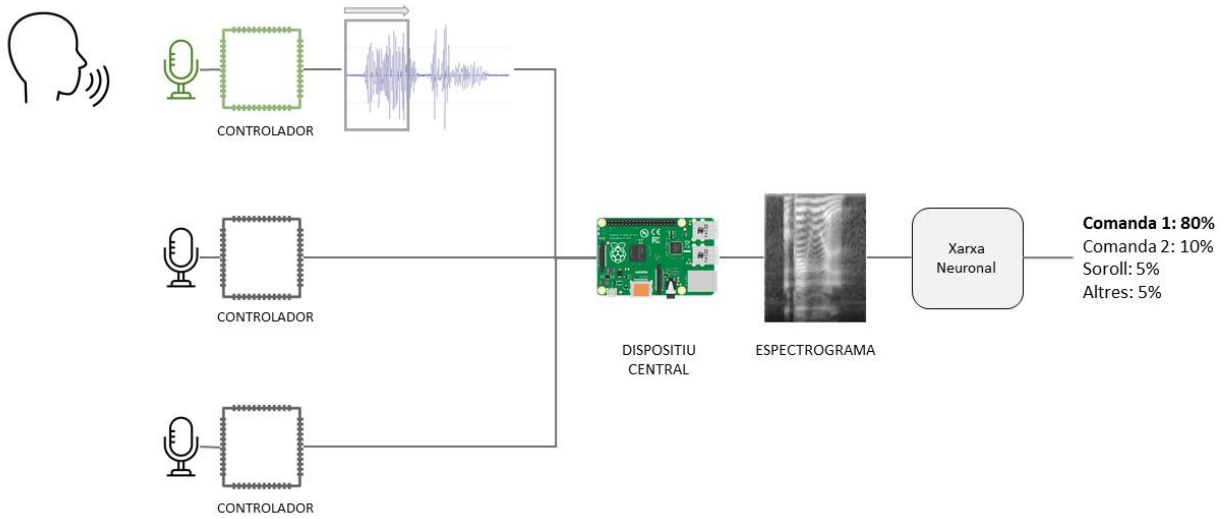

Figura 1 – Arquitectura del sistema de reconeixement de veu amb xarxa neuronal convolucional. El dispositiu controlador envia el senyal d'àudio al dispositiu central, el qual genera l'espectrograma per tal que la xarxa neuronal el classifiqui.

3.2. Reptes resolts i resultats obtinguts

Distinció entre veu humana i soroll de fons per extracció de propietats de les ones sonores

Per tal de no saturar el dispositiu central i mantenir la privacitat el màxim possible, els dispositius controladors juguen un paper important detectant la presència o absència de veu humana. En aquest procés no es veu involucrat cap algoritme d'intel·ligència artificial, sinó que s'utilitzen mètodes clàssics de detecció d'activitat de veu. En una primera etapa, es duen a terme diverses accions de filtratge, com la reducció del soroll de fons i l'amplificació del senyal. Després es pren una secció del senyal per extreure'n algunes característiques com el "zero-crossing rate" (ZCR), corresponent a la freqüència amb la qual el senyal passa de zero a un valor positiu o a un valor negatiu. Finalment s'aplica un algoritme de classificació per determinar si el senyal correspon o no a una mostra de veu. En cas afirmatiu, es notifica al dispositiu controlador per tal d'iniciar el reconeixement de veu.

Reconeixement de paraules clau per CNN

El reconeixement de paraules clau és el punt principal de l'aplicació. Permet a l'usuari interactuar amb el sistema i permet als micròfons actuar com un sensor sintètic, afegint funcionalitats d'intel·ligència ambiental. Per a la prova de concepte, s'ha aconseguit un sistema capaç de detectar dues paraules clau o "keywords": "Ok System" i "Ok Google". Quan el sistema detecta la primera keyword, s'espera a que l'usuari doni una de les comandes personalitzades. En el segon cas, el sistema entén que es tracta d'una comanda dirigida a Google. En aquest cas, si l'usuari així ho desitja, el sistema por enviar tot l'àudio capturat després de la comanda a la API de Google per tal de fer-ne el reconeixement de veu.

Reconeixement de comandes

Actualment, el reconeixement de comandes es realitza mitjançant la detecció de paraules clau associades a les comandes. En un futur, aquest procés s'ha d'actualitzar de cara a fer un sistema més flexible i obert, capaç de detectar paraules de manera genèrica i associar aquestes deteccions a les diferents comandes del sistema. Actualment, el sistema detecta 8 comandes personalitzades destinades a encendre i apagar dispositius: “turn” + “on/off” + un número de l’1 al 4.

3.3. Limitacions actuals de la tecnologia

Actualment la potència de càlcul dels dispositius de baix cost en el edge és força limitada i no és possible realitzar una detecció i processat de la veu de manera generalitzada. Per ara, la solució es limita a detectar paraules clau i comandes de veu predefinides per les quals la xarxa neuronal s’ha entrenat específicament. Per introduir noves comandes s’ha de tornar a entrenar la xarxa neuronal. En les properes tasques de recerca es volen focalitzar en fer el sistema mes escalable i flexible en la detecció de comanes de veu. També es vol investigar l’ús de noves arquitectures de detecció de paraules clau i reconeixement del llenguatge que funcionin tant amb el senyal d’àudio de forma directa i no sigui necessari construir l’espectrograma.

4. Potencial impacte de la solució

- Dotar d'interfícies naturals, basats en la veu, a dispositius que per cost, mides, facilitat de manipulació o proximitat no puguin tenir una interfície prou adient per a la seva manipulació. Per exemple a una torradora es podria subsistir el boto d'ajust de temperatura per un comana i un missatge de veu. Una llum es pot controlar a distancia amb la veu i una cuina es pot encendre o apagar encara que tinguem les mans mullades o ocupades a en una altre tasca.
- Prevenció d'us no autoritzat. Els sistemes de veu es poden entrenar per acceptar comanes de qualsevol interlocutor però també es poden fer per reconèixer un usuari o un grup d'usuaris. Aquesta facilitat ofereix un mecanisme d'autenticació de forma que nomes l'usuari reconegut podrà enviar comanes al dispositiu. Això resulta molt interessant en sistemes d'accés (panys de porta) o en l'ús de dispositius perillosos per nens petits o persones grans.
- Detecció de situacions disruptives: conflictes en espais públics, sorolls en zones de silenci.

CIDAI Centre of Innovation for Data tech and Artificial Intelligence

www.cidai.eu