Software per a llegir els llaivs (Lip Reading)

Javier Alejandro Camacho Machaca

Resum — Aquest treball presenta el desenvolupament d'un software capaç de llegir els llavis mitjançant el reconeixement d'un rostre. Amb aquest programari es busca facilitar el treball per a aquelles persones amb dificultats visuals o auditives, o en altres casos per a persones que vulguin saber que diu altra persona en una multitud o si està molt lluny. [Explicació de com s'ha aconseguit desenvolupar el software (mètode, eines, investigacions etc...) i dir els seus resultats.]

Paraules clau—Paraules clau del projecte, màxim 2 línies. ....
....

Abstract—Versió en anglès del resum. ....
....
....
....
....
....
....
....
....

Index Terms—Versió en anglès de les paraules clau. ....
....

1 Introducció - Context del treball

És sabut per la majoria que les persones amb dificultats o discapacitats tant visuals com auditives o tenen més difícil a l'hora d'interactuar amb altres persones, ja sigui perquè la resta de la gent no sap llenguatge de signes o la persona que es vulgui comunicar, té dificultats per veure que li vol dir l'altra persona o directament no ho pot veure. També estan totes aquelles persones que han perdut la capacitat de comunicar-se o entendre als altres. Per exemple a les persones que pateixen càncer de gola, perden la capacitat de parlar com va ser el cas de Val Kilmer. O en altre cas com el de Bruce Willis, que pateix d'Afàsia, que és una malaltia que provoca la pèrdua de la capacitat d'expressar o comprendre el llenguatge parlat o escrit, com a resultat del dany a les àrees del cervell que controlen el llenguatge, i així ens podem estendre a molts tipus de malalties o accidents sobtats que canvien la vida de les persones per sempre.

Ja hi ha diverses empreses o centres que han abordat aquest problema, però nosaltres volem aportar el nostre esforç a aquesta causa.

- *E-mail de contacte: 1566088@uab.cat*
- *Menció realitzada: Enginyeria de Computació*
- *Treball tutoritzat per: Coen Antens (Ciències de la Computació i Intel·ligència Artificial)*
- *Curs 2023/24*

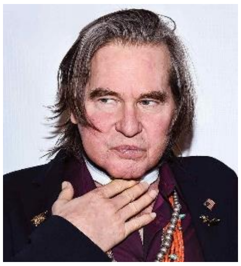

Fig. 1. Val Kilmer, en el 2015 es va sotmetre a una traqueotomia i avui porta sempre mocador al coll.

Per aquest motiu hem desenvolupat un software que serà capaç de llegir els llavis a través d'un reconeixement facial que se centrarà en les expressions dels llavis. I gràcies al (posar eina amb què s'aconsegueix aquest objectiu, per exemple, Deep Learning) obtindrem amb un vídeo dels llavis, al qual detectarem lletra per lletra el que vol dir, i per tant que paraules està dient l'altra persona.

Una vegada que tinguem transcrit l'input del vídeo a text gràcies al nostre software, tenim via lliure per estendre aquest resultat de diferents maneres, podríem utilitzar el nostre mateix software cap a un mateix, si nosaltres no tenim la capacitat de parlar, però sí de moure els llavis, per transcriure-ho a text i gràcies a una API "text-to-speech" expressar-ho amb un altaveu.

També podem fer servir aquesta API, possiblement treta de l'empresa OpenAI, per fer que persones que hagin perdut la capacitat de veure, puguin escoltar amb uns altaveus i el text transcrit pel nostre software, el que li vol dir una persona que s'està comunicant amb ell.

Altres casos importants de tenir en compte, fora de qualsevol classe de discapacitat, són aquells els quals hi ha molta gent i és molt difícil escoltar el que et vol dir el teu company, en aquest cas el nostre software no tindria cap problema en llegir els seus llavis. Un altre problema seria si et vols comunicar amb una persona que està molt llunyà, llavors no el podries escoltar ni tampoc veure'l clarament, llavors si s'aplica el nostre software a una càmera amb una resolució suficient, es podria saber que vol dir aquella persona. També hi ha casos fora de tota necessitat personal, com seria el cas dels comentaristes esportius, que gràcies a la nostra aplicació, tant les persones que no entenen el que diuen per què parlen força de pressa com les persones sordes, podran gaudir d'un partit.

Per comprendre el problema que planteja la barrera de l'idioma a l'hora de parlar, nosaltres ens centrarem en la fonètica dels llavis, ja que és similar en l'idioma llatí i el germànic. Dit això entrenarem el nostre model amb vídeos de persones parlant, i per això són preferibles vídeos de noticiaris, ja que les persones principals d'aquests vídeos pronuncien bé.

Objecius: (cambiar objetivos)

- Fer un algoritme per detectar la cara d'una persona
- Després de detectar la cara d'una persona, detectar i centrar-se en els moviments dels llavis
- Fer un software inicial que classifiqui quina vocal estàs dient.
- Buscar un Database amb vídeos de persones parlant amb el seu Ground Truth per a l'entrenament de l'algoritme.
- Fer un software que detecti, dins d'un rang, que paraules estàs dient de manera aïllada.
- Desenvolupar el software de les paraules perquè pugui classificar més paraules i de manera més seguida.

2 METODOLOGIA

L'organització i les eines utilitzades a l'hora de fer un projecte és molt important per treure'n el màxim rendiment al temps i tenir sempre una referència del que tenies planejat al que tens a la realitat. Per aquestes raons us explicaré quins són els que he decidit fer servir jo i el perquè.

2.1 Mètode Àgil

La metodologia que hem escollit és el Kanban, ja que en ser un projecte fet per una sola persona, no és necessari seguir unes pautes tan estrictes com ho seria si escollim Scrum. El mètode Kanban permet gestionar les tasques en tres grups, els "Per fer", "En curs" i "Fet". Aquesta metodologia és perfecta perquè permet gestionar les tasques d'una manera senzilla i efectiva.

Així que s'utilitzarà aquest mètode àgil per organitzar les tasques més globals i més importants d'una manera que sempre es vegi clarament en quant temps s'ha de fer i que subtasques el componen per assegurar-nos que es fa la tasca el més complet possible.

2.2 Eina de seguiment

Per fer el seguiment i la gestió de les tasques necessàries per fer aquest software, hem decidit utilitzar Jira Software. Aquesta eina, a part de ser una de les eines líders en aquest àmbit, compleix totes les necessitats dels meus requisits. Em permet tenir una interfície on puc aplicar el meu mètode Kanban, on gestiono les meves tasques en subtasques els quals puc ficar-los estats com "En progrés" o "Finalitzat" fora del sistema de Kanban. I també em permet crear un cronograma per posar fites de dates.

Les funcionalitats que s'han utilitzat del Jira Software són el "Tauler", que em permet veure les meves tasques com pòsits i puc moure'ls als estats del Kanban (Per fer, en curs i fet). Les "Incidències", on puc veure totes les meves fites i tasques per saber a més profunditat de què es tracten i on puc posar-los en "Finalitzat" o en "Fer". En últim lloc, està el "Cronograma", on puc veure el temps que s'ha de dedicar a cada fita i el temps que em queda per fer-los.

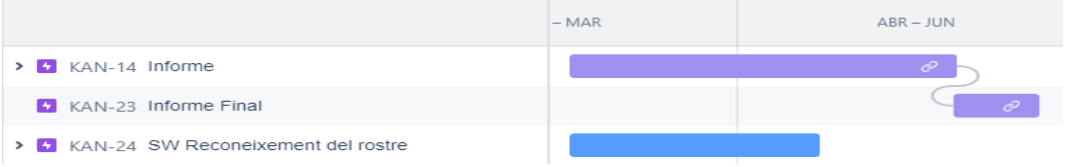

Fig. 2. Cronograma exportat de Jira Software

3 PLANIFICACIÓ (LISTA DE TAREAS)

Com és habitual, aquest projecte s'inicia amb una investigació i la recerca d'altres projectes o empreses que hagin realitzat alguna cosa similar, com ara la lectura de llavis. Aquestes són les referències que he pogut trobar... (Cal fer referència a la bibliografia amb tot el que estigui relacionat).

El primer gran repte o objectiu que es va proposar per a aquest projecte va ser desenvolupar un codi capaç de detectar els llavis d'una persona per determinar quina vocal està pronunciant. En primer lloc, vaig fer servir la llibreria 'mediapipe', específicament la funció 'face_mesh', per obtenir una malla del rostre. Un cop tenim tots els punts del rostre, ens centrem en els punts que formen els llavis per aconseguir un mapa de punts que utilitzarem per entrenar un model d'Aprenentatge Automàtic. Posteriorment, seguirem el mateix procés però fent servir un model de 'Deep Learning' per preparar-nos per a futurs reptes o objectius.

El següent pas és desenvolupar un software que em permeti identificar quina paraula individual està dient una persona. En un primer moment, aquest sistema es limitarà a un nombre fix de paraules, i si és viable, s'optimitzarà per aconseguir la màxima separació entre aquestes paraules quan es pronunciïn.

4 DESENVOLUPAMENT (EXPLICAR DESSARROLLO DE LAS TAREAS)

Com és habitual, no soc la primera persona que aspira a aconseguir l'objectiu de comprendre el que diu una persona sense la necessitat de poder escoltar-la. He pogut trobar diverses investigacions d'altres universitaris i empreses arreu del món que tenen el seu propi programari que assoleixen això. No obstant això, desitgem aportar el nostre esforç, ja que fer-ho sempre beneficia a tothom. Sigui per la varietat de processos o perquè és una iniciativa més actual, tot suma per avançar en el món de l'enginyeria.

Per tal de concloure el meu primer objectiu d'elaborar un programari que classifiqui vocals, he utilitzat la funció 'FaceMesh()' de la llibreria 'Mediapipe'. Aquesta funció, en primer lloc, detecta els rostres de les persones i després retorna una malla de punts del rostre expressada en coordenades i profunditat [x, y, z]. Ens concentrarem únicament en els 80 punts que formen els llavis per tal d'obtenir el gest en si mateix (Resultats d'aquesta funció en la imatge 'Fig. 3').

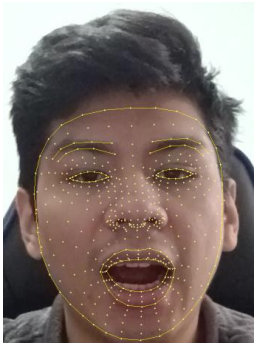

Fig. 3. Foto feta per mi mateix i processada per la funcio

Per obtenir un conjunt de dades adequat per a l'entrenament, afegiré un punt addicional que estarà situat al centre de la boca. Per aconseguir-ho, simplement calcularé la mitjana dels valors dels 80 punts existents, obtenint així el punt 81 també en format [x, y, z]. Per prescindir del format de llibreria dels punts i, així, poder dur a terme l'entrenament del model, substituiré les coordenades dels punts per la distància respecte al punt 81, que representa el centre de la boca. D'aquesta manera, en finalitzar aquest procés, disposarem només de 80 distàncies per a cada imatge que analitzem. A partir d'aquí, només queda entrenar el model. Per fer-ho més interessant, també utilitzarem aquest mateix procés de creació del conjunt de dades com a entrada per a un model de xarxes neuronals. Per poder visualitzar el resultat d'aquesta primera tasca, farem servir la llibreria 'Gradio', que mostrarà a la imatge 'Fig. 4'.

El darrer objectiu a complir abans de començar a predir paraules en lloc de lletres és trobar un conjunt de dades adequat per a nosaltres. Si no en podem trobar un que s'adapti a les nostres necessitats d'iniciar la predicció de paraules individuals, caldrà crear el nostre propi conjunt de dades, tal com vam fer per entrenar el classificador de vocals, ja que no existia cap conjunt d'imatge similar al que buscàvem.

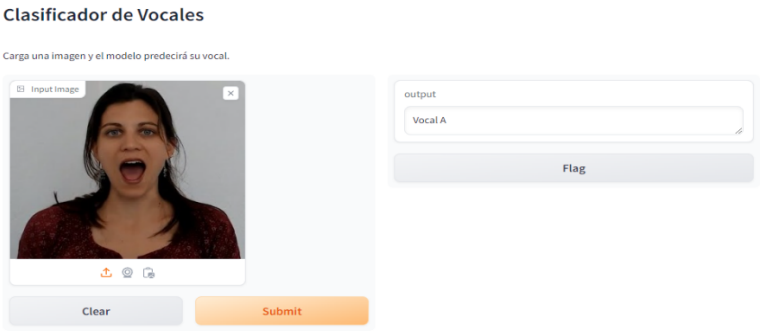

Fig. 4. Imatge extreta del resultat de Gradio amb el model de reconeixement de vocals exportat. Selecciona una imatge i obtindràs com a sortida quina vocal està pronunciant.

5 Conclusió

....

Agraïments

....

Bibliografia

[1] Yannis Assael, Brendan Shillingford, Shimon Whiteson and Nando de Freitas (2016, Novembre) LipNet [Online] URL: (https://arxiv.org/abs/1611.01599)

[2] Liopa (2015, Novembre) Deciphering speech from lip movements [Online] URL: (https://liopa.ai)

[3] Inclusive Interaction Lab (2021, Gener 12) LipType [Online] URL: (https://github.com/theiilab/LipType)

[4] Joseph Redmon (2016) YOLO [Online] URL: (https://www.v7labs.com/blog/yolo-object-detection)

[5] Meta Research (2022, Gener 6) AV-HuBERT [Online] URL: (https://github.com/facebookresearch/av_hubert)

[6] Pingcuhan Ma (2023, Juny 16) Auto-AVSR: Lip-Reading Sentences Project [Online] URL: (https://github.com/mpc001/auto_avsr)

[7] Pingcuhan Ma (2020, Juny 25) Lipreading using Temporal Convolutional Networks [Online] URL: (https://github.com/mpc001/Lipreading_using_Temporal_Convolutional_Networks)

[8] Pingcuhan Ma (2022, Març 1) Visual Speech Recognition for Multiple Languages [Online] URL: (https://github.com/mpc001/Visual_Speech_Recognition_for_Multiple_Languages)

[9] OMES (2021, Maig 26) Malla Facial (MediaPipe Face Mesh) | Python - MediaPipe – OpenCV [Online] URL: (https://www.youtube.com/watch?v=TCUi-pOXuCBQ&list=PLcz5q4ASTYQGKfBzjFBX5nVPkv-TUYn2Ro&index=3)

[10] Nachiketa Hebbar (2021, Abril 20) Image Classification Web App in Python| Keras +Gradio [Online] URL: (https://www.youtube.com/watch?v=aZ4wV4V_p9E)

[11] Spanish Pronunciation Academy (2017, Maig 24) Los sonidos del español: Vocal A [Online] URL: (https://www.youtube.com/watch?v=E7PKLMBAUtk)

Apèndix

A1. Secció d'apèndix

....

A2. Secció d'apèndix

....