الجمهورية العربية السورية
وزارة التعليم العالي والبحث العلمي
الجامعة الوطنية الخاصة
كلية الهندسة
قسم هندسة الحاسوب

مشروع تخرج بعنوان:

نظام تحويل لغة الشارة العربية إلى كلم منطوق

Arabic Sign Language to Speech Conversion System

دراسة أعدت لنيل درجة الإجازة في هندسة الحاسوب

إعداد:

.....

إشراف :

الدكتور .....

العام الدراسي 2025-2026

الإهداء

إلى اليد التي ترتفع في قاعة انتظار مزدحمة، تتحرك بثقة، فل يفهمها أحد.

إليـــك كتبنـــا هـــذا العمـــل. فكـــل نمـــوذج درّبنـــاه، وكـــل واجهـــة صمَّاها، وكل خط@ أصلحناه، كا? يدور حول سحال واحد: ماذا لو وقفت غدا في عيادة أو صيدلية أو دائرة رسمية، فوجدت م? يفهمك فورا؟ لم تك? بالنسبة إلينـــا «حالة استخدام» في وثيقة متطلبات؛ كنت أنت السبب الذي م? أجله تعلّمنا أ? الكلم قد يسك? اليدي?.

وإلى أهلنا، الذي? سمعوا منا ع? هذا المشروع أكثر مما سمعوا ع? أنفسنا، وصبروا على غيابنا خلف الشاشات والمقاطع والجداول، ثم كانوا أول م? فرح حي? نطـــق النظام جملته الأولى كاملة.

وإلى أساتذتنا في كلية الهندسة، الذي? علّمونا أ? الهندسة ليست أ? يعمل الشـــيء فحسب، بل أ? يعمل لم? يحتاجه، وأ? أصعب المشكلت ليست دائما أكثرها تعقيدا في الحساب.

وإلى زملئنا الذي? وقفوا معنا أمام الكاميرا يحدو? الإشارات ضاحكي? مخطـــئي? مصيبي?، فصارت أخطاحهم جزءا م? بيانات هذا العمل وم? ذاكرته.

نهديكم هذا الجهد، على أمل أ? تصمت فيه المسافة بي? يد تتكلم وأذ? تسمع.

بسم ال الرحم? الرحيم

{يرفعِ ال ُ الذي? امنوا منكمْ والذي? اوتُوا العلمَ درجات}

أما بعد...

الحمد ل أولاً وآخرا، الذي يسّر لنا طريق هذا العمل م? فكرة عابرة إلى نظام يعمل بي? أيدي الناس.

كلمة الشكر

نتقدم بجزيل الشكر إلى مشرف مشروعنا، الذي رافق العمل منذ كا? سحالاً مفتوحا: هل تستطيع الكاميرا أ? تفهم اليد؟ ووجّهنا في قرارات مفصلية، منها أ? نبدأ بمفردات المجال الصحي قبل التوسع، وأ? نقيس نجاحنا على أشخاص لم يرهم النموذج، لا على أرقام مريحة. كما نشكر كلية الهندسة في الجامعة الوطنية الخاصة وأساتذتها على ما وفّروه م? بيئة وتشجيع طوال سنوات الدراسة.

ونشكر أهلنا وأصدقاءنا، وكل م? أدّى أمام كاميراتنا إشارة، أو قرأ مسودة، أو سمع منا شرحا مكررا للمشكلة نفسها بصبر واهتمام. ونشكر أخيرا الباحثي? الذي? أتاحوا مجموعتي KArSL وArASL2018 للمجتمع العلمي؛ فلولا إتاحة عملهم لما وُجد هذا المشروع أصلً.

ملخص

يقدّم هذا المشروع نظام «إشارة إلى كلم»، وهو تطبيق ويب يحوّل لغة الإشـــارة العربية إلى جمل عربية منطوقة اعتمادا على كاميرا الهاتف وحده، دو? أي معالجة على خادم ودو? أ? يغادر الفيديو جهاز المستخدم. يعمل النظام بثلثـــة أوضـــاع متكاملة: وضع إشارات سهلة بيد واحدة يُقرأ بقواعد هندسية على نقاط اليد الإحدى والعشري? فل يحتاج إلى تدريب ويعمل مع أي مستخدم، ووضع حـــروف يهجّئ الكلمـــات بشـــبكة عصـــبية صـــغيرة م? ثلث طبقـــات درّبت على مجموعـــة ArASL2018 فبلغت دقتها 89.1%، ووضع كلمات يتعرف على إشارات متحركة كاملة بشبكة التفافية زمنية تعمل بصيغة ONNX داخل المتصفح على متجهـــات خصائص م? 378 عددا لكل إطار مستخرجة@دوات MediaPipe. درّبنا نموذج الكلمات على مجموعة KArSL بمرحلتي?: مئة إشـــارة صـــحية بمقـــاطع بلغت 4,800 مقطع فبلغت الدقة 85.8% بمنهجية استبعاد محدٍّ كامل، ثم 261 إشـــارة بمقاطع بلغت 13,286 مقطعا فبلغت الدقة 76.5% عنـــد أول تخمي? و92.0% ضم? أول خمسة تخمينات. وتُبنى الجملة النهائية بمحرك قوالب لغوية يعتمد فئات الإشارات، وتُنطق بمقاطع صوتية مولّدة مسبقا، ويقدّم التطبيق قاموسا مرئيا متكاملً وإمكانية تعليم إشارات جديدة محليا، ويعمل دو? اتصال بوصفه تطبيق ويب تقدميا. وتوثّق فصول هذا الكتاب معمارية النظام وخطي التدريب والتشغيل ونتائجه وحدوده الحالية وخطوات تطويره.

كلمات مفتاحية: لغة الشارة العربية، التع "رف على الش 'ارات، نق 'اط المعلم، MediaPipe ، التعلم العميق، تحويل النص إلى كلم، تطبيق ويب تقدمي.

Abstract

This project presents Sign-to-Speech, a web application that turns Arabic Sign Language into spoken Arabic sentences using only a phone camera. All processing runs inside the browser: no server ever receives the video, and the application keeps working offline after its first load. The system offers three modes. An easy-gesture mode reads nine one-hand shapes with geometric rules over the 21 hand landmarks, so it needs no training and works for any user. A letters mode spells words through a small three-layer network trained on the ArASL2018 dataset, reaching 89.1% accuracy. A words mode recognizes full dynamic signs with a temporal convolutional network exported to ONNX, fed by 378-number feature vectors per frame extracted with MediaPipe. The words model was trained on the KArSL database in two stages: 100 health signs with 4,800 clips scored 85.8% under leave-one-signer-out evaluation, then 261 signs with 13,286 clips scored 76.5% top-1 and 92.0% top-5 accuracy. A rule-based template engine assembles the recognized signs into Arabic sentences, which are spoken through pre-generated audio clips. The application also ships a visual dictionary of its whole vocabulary and lets users teach it new signs locally on their own devices.

Keywords: Arabic Sign Language, sign recognition, landmarks, MediaPipe, deep learning, text-to-speech, progressive web app.

جدول الأشكال.....VIII

جدول الجداول.....X

جدول المصطلحات.....XI

جدول التعاريف.....XIII

الفصل الأول: مقدمة عامة.....1

1-1 مقدمة.....2

1-2 أهمية المشروع.....4

1-3 الأهداف.....5

1-4 الدراسات السابقة.....6

1-5 المتطلبات الوظيفية.....8

1-6 المتطلبات غير الوظيفية.....9

الفصل الثاني: الدراسة النظرية.....11

2-1 لغة الإشارة العربية وخصائصها.....12

2-2 الرحية الحاسوبية ونقاط المعلم.....13

2-3 مكتبة MediaPipe ونماذج التتبع.....14

2-4 تصنيف التسلسلت الزمنية للشارات.....15

2-5 مجموعات البيانات التدريبية.....16

2-6 م? الإشارة إلى الكلم: الصياغة والنطق.....17

الفصل الثالث: الأدوات والتقنيات المستخدمة.....19

3-1 بيئة التطوير ولغة JavaScript وأداة Vite.....20

3-2 مكتبة MediaPipe Tasks Vision.....20

3-3 صيغة ONNX ومشغّل ONNX Runtime Web.....21

3-4 بيئة التدريب: Python ومكتبة PyTorch.....22

3-5 توليد الصوت: مـ? Piper إلى المقاطع الجاهزة.....23

3-6 النشر: Cloudflare Pages وتطبيق الويب التقدمي.....23

الفصل الرابع: التطبيق العملي.....25

4-1 مقدمة.....26

4-2 معمارية النظام وخط المعالجة.....26

4-3 وضع الإشارات السهلة.....27

4-4 وضع الحروف.....30

4-5 وضع الكلمات وخط التدريب.....32

4-6 صياغة الجملة والنطق.....35

4-7 القاموس والتدريب الذاتي.....36

4-8 الاختبار والنتائج.....37

4-9 الصعوبات وحدود النظام.....41

الخاتمة والآفاق المستقبلية.....42

الملحق.....45

المراجع.....47

الشكل (1-1): الوضاع الثلثة في نظام إشارة إلى كلم: الشارات السهلة، والحروف، والكلمات.-.....3
الشكل (1-2): الواجهة الرئيسة للنظام على الهاتف قبل بدء التعرف.-.....6
الشكل (1-2): مخطط مبسط لنقاط معلم اليد الواحدة وعددها 21 نقطة.-...14
الشكل (1-4): خط معالجة وضع الكلمات من الكاميرا حتى إضافة الشارة إلى الجملة.-.....27
الشكل (2-4): تعرف النظام على إشارة «أحبك» في وضع الشارات السهلة أثناء اختبار حي بكاميرا وهمية.-.....29
الشكل (3-4): دليل حروف البجدية الشارية داخل التطبيق بصور حقيقية وزر تجربة لكل حرف.-.....31
الشكل (4-4): صور مرجعية من مجموعة ArASL2018 لول ثمانية حروف في دليل التطبيق.-.....32
الشكل (5-4): خط تدريب نموذج الكلمات من مقاطع البيانات حتى النشر داخل التطبيق.-.....33
الشكل (6-4): توزع مفردات النموذج الحالي وعددها 261 إشارة على فئاتها.-.....34
الشكل (7-4): مثال صياغة جملة من إشارتين بمحرك القوالب.-.....35
الشكل (8-4): شاشة قاموس الشارات مع رسوم الحروف المتحركة وفئات المفردات.-.....36
الشكل (9-4): الواجهة الرئيسة للنظام على الحاسوب مع لوحة الجملة وأزرار التحكم.-.....37
الشكل (10-4): مقارنة دقة النموذجين باستبعاد مؤدٍ كامل عبر المؤدين الثلثة.-.....39

الشكل (11-4): لقطتا اختبار حي: تعرف النظام على إشارتي «تمام» و«اتصل بي» من صور مولدة عبر كاميرا وهمية.-.....40

جدول الجداول

الجدول (1-1): ملخص الدراسات السابقة وموقع نظامنا بينها.-.....7
الجدول (1-2): مكونات متجه الخصائص لكل إطار وعددها 378 عددا.-...16
الجدول (2-2): مجموعتا البيانات المستعملتان في تدريب نماذج النظام.-...17
الجدول (1-3): الدوات والتقنيات المستعملة في بناء النظام وتشغيله.-....22
الجدول (1-4): مقارنة بين أوضاع التعرف الثلثة في النظام.-.....34
الجدول (2-4): نتائج نموذج المئة إشارة باستبعاد مؤلف كامل.-.....38
الجدول (3-4): نتائج نموذج 261 إشارة باستبعاد مؤلف كامل.-.....38
الجدول (4-4): نتائج اختبار حي مستقل لوضع الشارات السهلة بكاميرا وهمية وصور مولدة.-.....40

جدول المصطلحات

المصطلح	الصيغة الكاملة	المعنى بالعربية
API	Application Programming Interface	واجهة برمجة التطبيقات
ArASL2018	Arabic Alphabets Sign Language Dataset	مجموعة صور حروف لغة الإشارة العربية
CC BY 4.0	Creative Commons Attribution 4.0	رخصة المشاع الإبداعي مع وجوب النسبة
Cloudflare Pages	Cloudflare Pages Hosting Platform	منصة نشر المواقع الثابتة المستعملة في إطلق التطبيق
ElevenLabs	ElevenLabs Speech Platform	خدمة توليد الأصوات المستعملة لإنتاج المقاطع
getUserMedia	MediaDevices.getUserMedia API	واجهة المتصفح القياسية للوصول إلى الكاميرا
Hand Landmarker	MediaPipe Hand Landmarker Task	مهمة تتبع اليد واستخراج نقاطها في MediaPipe
Holistic	MediaPipe Holistic Landmarker	الحل الشامل لتتبع اليد؟ والوضعية والوجه معا
HTTPS	Hypertext Transfer Protocol Secure	بروتوكول النقل الآمن؟، وهو شرط عمل الكاميرا في المتصفح
KArSL	Arabic Sign Language Database (KFUPM)	قاعدة بيانات لغة الإشارة العربية م؟ جامعة الملك فهد
MediaPipe	MediaPipe Framework (Google)	إطار بناء خطوط الإدراك العاملة على الجهاز
NumPy	Numerical Python Library	مكتبة الحوسبة العددية في بيئة التدريب
ONNX	Open Neural Network Exchange	صيغة التبادل المفتوحة لتمثيل نماذج التعلم الآلي
ONNX Runtime Web	ONNX Runtime for Web Browsers	مشغّل نماذج ONNX داخل المتصفح
Piper	Piper Text-to-Speech Engine	محرك تحويل نص إلى كلم مفتوح المصدر يعمل محليا
PWA	Progressive Web App	تطبيق ويب تقدّمي قابل للتثبيت ويعمل دو؟ اتصال
PyTorch	PyTorch Deep Learning Library	مكتبة التعلم العميق المستعملة في التدريب
Service Worker	Service Worker API	عامل الخدمة الذي يخزّ؟ التطبيق للعمل دو؟ اتصال
SpeechSynthesis	Web Speech API —	واجهة نطق النص@صوات الجهاز في

المصطلح	الصيغة الكاملة	المعنى بالعربية
	SpeechSynthesis	المتصفح
Tasks Vision	MediaPipe Tasks Vision Library	حزمة مهام الرحية م? MediaPipe للمتصفح
Top-1 / Top-5	Top-1 / Top-5 Accuracy	الدقة عند أول تخمي? / ضم? أول خمسة تخمينات
Vite	Vite Build Tool	أداة بناء حزم الويب وخادم التطوير
WebAssembly	WebAssembly (Wasm)	صيغة تنفيذ سريعة داخل المتصفح تشغّل النماذج
Whisper	Whisper Speech Recognition Model	نموذج تحويل الكلم إلى نص استُعمل لفحص المقاطع الصوتية
Wrangler	Cloudflare Wrangler CLI	أداة سطر الأوامر للنشر إلى Cloudflare

جدول التعاريف

المصطلح	التعريف
استبعاد مجدٍّ كامل	طريقة تقييم يُحتجَز فيها مجدٍّ كامل م? التدريب ويُختبر النموذج على أدائه وحده، لقياس التعميم على أشخاص جدد.
الاستدلال المحلي	تشغيل النموذج وإنتاج التنبؤ على جهاز المستخدم نفسه دو? إرسال بيانات إلى خادم.
الإشارة المعزولة	إشارة واحدة ذات بداية ونهاية واضحتي? تحدّى منفردة، كما في مجموعات البيانات المستعملة في التدريب.
بوابة الثقة	قاعدة قرار تطبق على احتمال النموذج: قبول ونطق مباشر فوق عتبة عليا، وعرض خيارات بي? عتبتي?، ورفض دو? العتبة الدنيا.
التقطيع	تحديد بداية الإشارة ونهايتها داخل بث الكاميرا المستمر، وفي نظامنا يتم بخفض اليدي? ع? خط الراحة.
التهجئة بالأصابع	تمثيل الكلمات حرفا حرفا@ شكال يد ثابتة، وتستعمل للسماء والكلمات غير الموجودة في القاموس.
زم? التثبيت	المدة التي يجب أ? يبقى فيها شكل اليد ثابتا قبل أ? يسجّله النظام إشارة مجكدة، ودرءا للتكرار العرضي.
شكل اليد	هيئة الأصابع والكف في لحظة ما، وهو أحد معالم الإشارة التي تُبنى عليها القواعد والتصنيف.
الصياغة بالقوالب	بناء الجملة م? فئات الإشارات المتعرف عليها بقواعد لغوية مكتوبة صراحة، بدل نموذج لغوي عام.
عتبة الثقة	قيمة احتمال فاصلة يُبنى عليها قرار قبول نتيجة التعرف أو عرض بدائلها أو رفضها.
متجه الخصائص	تمثيل عددي مضغوط لإطار أو لإشارة، يستعمله النموذج مدخلً للتصنيف، وفي نظامنا 378 عددا لكل إطار.
نقطة المعلم	إحداثيات نقطة مفصلية محددة في اليد أو الجسم أو الوجه، يستخرجها المتعقّب م? كل إطار فيديو.

الفصل الول:

مقدمة عامة

1-1 مقدمة

في قاعة انتظار عيادة مزدحمة يقف شاب أمام موظف الاستقبال، يرفع يديه ويحرّكهما بترتيب مدروس: كفّ تشير إلى الرأس، ثم أصابع تنقبض قرب الصدغ. هو يقول بلغة الإشارة: «عندي صداع». الموظف لا يعرف هذه اللغة، فيبتسم مرتبكا ويشير إلى مقعد الانتظار. تمرّ الدقائق، ويتكرر المشهد نفسه عند شباك صيدلية، وفي دائرة حكومية، وعلى باب مشفى. المشكلة ليست أ? الشاب لا يستطيع الكلم؛ مشكلته أ? م? حوله لا يفهمو? لغته، وأ? مترجم لغة الإشارة غير موجود في معظم هذه المواقف اليومية، لأ? المترجمي? قلّة، والحاجة إليهم لحظية لا تنتظر موعدا [1].

حجم هذه الفجوة أكبر مما نظ?. فبحسب التقرير العالمي للسمع الصادر ع? منظمة الصحة العالمية عام 2021، يعيش أكثر م? مليار ونصف المليار إنسا? بدرجة ما م? فقدا? السمع، ويحتاج 430 مليونا منهم إلى@هيل لمعالجة فقدا? سمع مُعطِّل، بينهم 34 مليو? طفل، ويعيش نحو 80% م? هـ<لاء في البلدا? المنخفضة والمتوسطة الدخل حيث خدمات الترجمة و@هيل أصلً شحيحة. وتتوقع المنظمة أ? يقترب عدد م? يعانو? درجة م? فقدا? السمع م? مليارى? ونصف المليار بحلول عام 2050، وأ? تتجاوز كلفة فقدا? السمع غير المعالج 980 مليار دولار سنويا على الاقتصاد العالمي [1].

أما لغة الإشارة نفسها فهي لغة كاملة لا مجموعة إيماءات عابرة. يقدّر الاتحاد العالمي للصم أ? عدد الصم في العالم يتجاوز 70 مليونا، يتحدثو? أكثر م? 300 لغة إشارة مختلفة، وقد اعترفت اتفاقية حقوق الأشخاص ذوي الإعاقة بلغات الإشارة بوصفها لغات مساوية للغات المنطوقة، وخصّصت الجمعية العامة للمم المتحدة يوم 23 أيلول يوما دوليا للغات الإشارة [2]، [3]. وفي العالم العربي تزداد الصورة تعقيدا: لكل بلد تقريبا لهجته الإشارية الخاصة، والموارد الرقمية المتاحة للغة الإشارة

العربية — م؟ قواميس ومجموعات بيانات وأدوات ترجمة — أقل بكثير م؟ مثيلتها الإنكليزية، وهذا بالضبط ما يجعل أي نظام عملي في هذا المجال مساهمة حقيقية لا تمرينا برمجيا [2].

م؟ هنا وُلد مشروعنا: نظام «إشارة إلى كلم»، وهو تطبيق ويب يفتح كاميرا الهاتف، يقرأ إشارات المستخدم الأصم، ويحوّلها إلى جملة عربية منطوقة بصوت مسموع يفهمه م؟ يقف أمامه. اخترنا أ؟ يعمل كل شيء داخل المتصفح نفسه: لا خادم يعالج الصور، ولا نموذج ذكاء اصطناعي مستضاف يتلقى الفيديو، فالفيديو لا يغادر جهاز المستخدم أصلً. ولأ؟ مستخدمنا قد يقف في مكا؟ بل تغطية، صمّمنا التطبيق ليعمل دو؟ اتصال بعد تحميله أول مرة. ويبيد؟ الشكل (1–1) الأوضاع الثلثة التي يقدمها النظام للمستخدم، وهي أوضاع صُممت لتغطي مستويات مختلفة م؟ المهارة والحاجة، وسنشرحها بالتفصيل في الفصل الرابع.



الشكل (1–1) : الوضاع الثلثة في نظام إشارة إلى كلم: الشارات السهلة، والحروف، والكلمات.

نلحظ في الشكل أ؟ الأوضاع الثلثة ليست بدائل متنافسة بل درجات على سلم واحد: وضع الإشارات السهلة يعمل بقواعد ثابتة على شكل اليد ولا يحتاج تدريبا، ووضع الحروف يهجّئ الكلمات حرفا حرفا بشبكة عصبية صغيرة، ووضع الكلمات يتعرف على إشارات كاملة متحركة بنموذج مدرّب على مئات الإشارات. هذا التدرج كا؟

قرارا مقصودا: المستخدم الجديد يبدأ بما ينجح معه فورا، وم? يحتاج مفردات أوسع ينتقل إلى الأوضاع الأقوى دو ? أ? يتعلم واجهة جديدة.

2-1أهمية المشروع

تنبع أهمية هذا المشروع م? أنه يعالج لحظة تواصل حقيقية تتكرر كل يوم: لحظة يقف فيها إنسا? أصم أمام إنسا? يسمع، في عيادة أو صيدلية أو دائرة رسمية أو متجر، ويحتاج أ? يقول جملة واحدة قصيرة: «عندي ألم في صدري»، «أحتاج هذا الدواء»، «أريد الذهاب إلى المشفى». هذه الجمل القصيرة هي تحديدا ما صُمم نظامنا ليقوله، وقد اخترنا مفردات النموذج الأول م? المجال الصحي لهذا السبب: لأ? موقف العيادة هو أكثر موقف تكو ? فيه الترجمة غائبة والرسالة مستعجلة في آ? واحد.

وللمشروع أهمية تقنية لا تقل ع? أهميته الإنسانية. فمعظم أنظمة التعرف على لغة الإشارة المنشورة في الأبحاث تعمل على حواسيب قوية أو خوادم سحابية، وتُعرض نتائجها في مختبرات لا في أيدي الناس. أما نظامنا فيعمل كاملً على هاتف عادي: نماذج صغيرة الحجم، واستدلال محلي، وواجهة عربية مصممة لهاتف يُمسك بيد واحدة بينما اليد الأخرى تحدي الإشارة. إثبات أ? هذا ممكن بدقة عملية — 85.8% في نموذج المئة إشارة و76.5% في نموذج 261 إشارة على حدٍّ لم يره النموذج م? قبل — يفتح الباب أمام تطبيقات مشابهة في التعليم والخدمات الحكومية وخدمة العملاء، دو ? كلفة خوادم ودو ? التضحية بخصوصية المستخدم.

وهناك أهمية ثالثة تتعلق بالبيانات العربية. أثناء بناء المشروع اضطررنا إلى التعامل مع شحّ الموارد: مجموعة بيانات عربية واحدة رئيسية للشارات الكاملة، وأخرى لصور الحروف، وكلتاهما محدودتا عدد المحدين. توثيق هذه التجربة بصراحة — بما فيها الأخطاء التي اكتشفناها مثل انعكاس اتجاه الكاميرا في بيانات التدريب — يختصر الطريق على أي فريق عربي يبني نظاما مشابها بعدنا.

1-3 الهداف

هدفنا العام كا? واحدا وواضحا: أ? يستطيع الإنسا? الأصم أ? يقول جملة عربية مفهومة بصوت مسموع، بهاتفه وحده، في المكا? الذي يقف فيه، دو ? وسيط بشري ودو ? اتصال بالإنترنت. وكل قرار تقني في المشروع رُدّ إلى هذا الهـــدف: حي? تعارضت الدقة مع العمل على الجهاز اخترنا نموذجا أصغر يعمل على الجهـــاز، وحي? تعارضت كثرة المفردات مع بساطة الاستعمال قدّمنا وضعا سهلً يعمل فورا بل تدريب.

وم? هذا الهدف تفرعت أهداف عملية: بناء خط معالجة كامـــل م? الكـــاميرا إلى الصوت داخل المتصفح؛ وتدريب نموذج تعرف على إشارات عربية كاملة يعمّم على أشخاص لم يتدرب عليهم، وقيس ذلك بمنهجية صارمة هي استبعاد محدٍّ كامل م? التدريب ثم اختباره؛ وتقديم وضع حروف يتيح تهجئة الأســـماء والكلمـــات غـــير الموجودة في القاموس، لأ? أي قاموس مهما كبر سيعجز ع? اسم شخص أو دواء؛ وبناء قاموس مرئي داخل التطبيق يعلّم المستخدم الإشارات نفسها برسوم متحركة مستخرجة م? البيانات الحقيقية، حتى لا يكو ? التطبيق أداة استعمال فقط بل أداة تعلّم أيضا.

وأردنا أيضا أ? يبقى النظام قابلً للنمو بيد مســـتخدمه: يســـتطيع أي شـــخص أ? يعلِّم التطبيق إشارة جديدة بكاميرته في ثوا?، وأ? يغيّر معنى أي إشارة سهلة ليناسب لهجته المحلية، لأ? لهجات الإشارة تختلف م? بلد لآخر ولا يستطيع نموذج واحد مدرّب مركزيا أ? يغطيها كلها. ويبي? الشكل (2-1) الواجهة الرئيسة للنظام كمـــا تظهر على الهاتف، وهي واجهة مقصودة البساطة: زر بدء واحد كبير، ولوحـــــة تعرض الجملة وهي تُبنى.



الشكل (2–1): الواجهة الرئيسة للنظام على الهاتف قبل بدء التعرف.

نلحظ في الواجهة غياب أي إعدادات معقدة ع؟ الشاشة الأولى: الكاميرا والجملة وزر البدء فقط. صُمم هذا ع؟ قصد، لأ؟ مستخدمنا يستعمل التطبيق وهو واقف أمام شخص آخر ينتظر؛ فكل ثانية يقضيها في القوائم هي ثانية صمت محـــرج. أمـــا القاموس والإعدادات فموجودة لكنها على بعد لمسة واحدة، لا تعترض طريق المهمة الأساسية.

4–1الدراسات السابقة

اعتمدنا في بناء نظامنا على ركيزتي؟ بحثيتي؟ رئيسيتي؟، هما مجموعتا البيانـــات العربيتا؟ المتاحتا؟ للباحثي؟، وما بُني عليهما م؟ أعمال. الركيزة الأولى مجموعة KArSL للغة الإشارة العربية، وهي قاعدة بيانات سجّلها باحثو جامعة الملك فهـد للبترول والمعاد؟ بكاميرا Kinect V2، وتضم 502 إشارة معزولة أدّاها ثلثـــة

مدي? محترفي? بخمسي? تكرارا لكل إشارة، أي 75,300 عينة في المجموع. وقد استخدمنا منها في مرحلتنا الأولى الإشارات م? 0071 إلى 0170، وهي مئة إشارة معظمها م? المجال الصحي: أعراض وأعضاء وأدوية وفحوص وأماك?، ثم وسّعنا المفردات إلى 261 إشارة في المرحلة الثانية [4].

والركيزة الثانية مجموعة ArASL2018 لصور حروف لغة الإشارة العربية، وتضم 54,049 صورة لاثنتي? وثلثي? فئة تمثل حروف الأبجدية وأشكالاً قريبة منهـــا، شارك في جمعها أربعو? شخصا، وهي متاحة برخصة المشاع الإبداعي CC BY 4.0. م? هذه المجموعة درّبنا شبكة وضع الحروف، واستعرنا صورها المرجعية لعرض شكل كل حرف داخل التطبيق، بعد أ? اكتشفنا بالتجربة أ? الرسوم الهيكلية للحروف لا يفهمها المستخدمو? [5].

وعلى صعيد طرق التعرف، درسنا أعمالاً حاولت حل المشكلة نفســـها@ســـاليب مختلفة. م? أبرزها عمل لُقما? على مجموعة KArSL نفسها بشبكة ذات تيار? (Two-Stream) تستفيد م? حركة الفيديو المتراكمة، وقد حقق نتـــائج قويـــة على المفردات الكاملة للمجموعة وبيّ? أ? معلومات الحركة وحدها تحمل جزءا كبيرا م? معنى الإشارة [6]. ودرسنا أيضا الاتجاه السائد في الأنظمة الحديثة القائم على نقاط المعلم (Landmarks): بدل إدخال صورة الفيديو الخام إلى الشبكة، تُستخرج أولاً إحداثيات نقاط اليدي? والجسم والوجه ثم يُصنّف تسلسل هذه النقاط، وهو الاتجاه الذي تبنّاه نظامنا لأنه يصغّر حجم النموذج بمرتبات ويجعله يعمل على الهاتف. ويلخص الجدول (1–1) أبرز الدراسات التي انطلقنا منها وموقع نظامنا بينها.

الجدول (1–1): ملخص الدراسات السابقة وموقع نظامنا بينها.

الدراسة	البيانات	الطريقة	أبرز نتيجة
قاعدة KArSL [4]	502 إشارة، 3 مدي?، Kinect	تسجيل ووصف القاعدة	75,300 عينة مرجعية للباحثي?
مجموعـــــــــــة	54,049 صـــورة، 32	جمع وتصنيف صـــور	قاعدة الحروف العربيـــة

الدراسة	البيانات	الطريقة	أبرز نتيجة
ArASL2018 [5]	حرفا	ثابتة	الإشارية الأكثر استعمالاً
لقما؟ 2022 [6]	KArSL-502	شبكة ذات تيارية؟ مـــع حركة متراكمة	دقـــة مرتفعـــة على المفردات الكاملة
نظامنا	KArSL + ArASL2018	نقاط معلم وشبكة التفافية زمنية على الجهاز	85.8% على 100 إشـــارة و76.5% على 261 إشارة لمحدٍّ جديد

نلحظ م؟ الجدول أ؟ الدراسات السابقة أجابت ع؟ سحال «هل يمكـ؟ التعرف على الإشارة؟» داخل المختبر، بينما سحال مشروعنا مختلف: «هل يمكـ؟ أ؟ يعمل هذا على هاتف مستخدم حقيقي، دو؟ خادم، وبلغته؟». لهذا لم ننافس تلك الأعمال على الدقة المطلقة، بل بنينا عليها: استعملنا بياناتها نفسها، وقارنّا بمنهجية التقييم الصارمة ذاتها (استبعاد محدٍّ كامل)، ثم نقلنا النموذج إلى بيئة التشغيل الحقيقية وهي المتصفح.

5-1المتطلبات الوظيفية

حددنا المتطلبات الوظيفية للنظام انطلقا م؟ سيناريو الاستعمال الأساسي: مستخدم أصم يقف أمام شخص يسمع ويريد إيصال رسالة. وتتلخص هذه المتطلبات في ما @تي:

- التقاط الفيديو م؟ كاميرا الهاتف الأمامية أو الخلفية، واستخراج نقـــاط معلم اليدي؟ والجسم والوجه م؟ كل إطار لحظيا.
- التعرف على الإشارات بثلثة أوضاع: إشارات سهلة بيد واحدة تعمل بقواعد ثابتة، وحروف أبجدية بالتهجئة، وكلمات كاملة متحركة بنموذج عصبي.
- بناء جملة عربية سليمة الترتيب م؟ الإشارات المتعرف عليها بقواعد لغوية تعتمد تصنيف كل إشارة، ونطقها بصوت عربي مسموع.
- عتبات ثقة متدرجة: تُضاف الإشارة الواثقة مباشرة، وتُعرض خيارات عند الثقة المتوسطة، ويُطلب م؟ المستخدم إعادة الإشارة عند انخفاض الثقة.

قاموس مرئي يعرض كل إشارة في المفردات برسم متحرك مستخرج م? بيانات حقيقية، مع بحث وتصنيفات ووضع تدريب يقيّم أداء المستخدم دو? إضافة نتيجته إلى الجملة.

- إمكانية تعليم التطبيق إشارات جديدة محليا وإعادة تعيي? معاني الإشارات السهلة، وحفظ ذلك على الجهاز نفسه.
- العمل دو? اتصال بعد التحميل الأول، بوصفه تطبيق ويب تقدميا قابلً للتثبيت على الشاشة الرئيسة.

6-1المتطلبات غير الوظيفية

أما المتطلبات غير الوظيفية فقد فرضتها طبيعة الاستعمال اليومي والموقف الاجتماعي الحساس الذي يعمل فيه النظام، وهي:

- الخصوصية: لا يغادر الفيديو جهاز المستخدم إطلقا؛ كل المعالجة محلية، وما يُحفظ على الجهاز هو نقاط معلم رقمية لا صور.
- زم? الاستجابة: يجب أ? يشعر المستخدم أ? التعرف فوري؛ لذا قِيس أداء الاستدلال على وحدة المعالجة المركزية وكا? بحدود 1.5 ميلي ثانية للشارة الواحدة في النموذج العصبي، والزم? الفعلي تحكمه الكاميرا واستخراج النقاط لا التصنيف.
- الحجم: مجموع ما يُحمّل م? نماذج محدود ببضعة ميغابايتات (نموذج الكلمات نحو 3 ميغابايتات، ونموذج تتبع اليد 7.8 ميغابايتات، وشبكة الحروف 353 كيلوبايتا) ليبقى التحميل الأول ممكنا على اتصالات ضعيفة.
- سهولة الاستعمال: واجهة عربية كاملة، أزرار كبيرة، وعدد خطوات قليل حتى أول جملة منطوقة.

قابلية التوسع: إضافة مفردات جديدة تتم بإعادة التدريب بالأدوات نفسها دو ? تغيير بنية التطبيق، لأ? ملفي استخراج الخصائص في التدريب والمتصفح متطابقا? بالتصميم.

الفصل الثاني:

الدراسة النظرية

2-1
لغة الشارة العربية وخصائصها

لغة الإشارة ليست ترجمة حرفية للكلم المنطوق ولا تهجئة مستمرة بالأصابع؛ إنها لغة قائمة بذاتها لها نحوها وترتيبها الخاص. ففي جملة منطوقة نقول «أنا عنـــدي صداع وأحتاج دواء»، أما في الإشارة فيحدي المستخدم عادة إشارات مفردة متتابعة: «صداع» ثم «دواء»، وتقع مهمة وصلها ببعضها في جملة سليمة على نظام مثـــل نظامنا. ولهذا السبب لم يك? التعرف على الإشارة إلا نصف المشكلة؛ النصف الآخر هو صياغة الجملة، وقد عالجناه بقوالب لغوية تعتمد تصنيف كل إشـــارة (فعـــل، عَرَض، حاجة، مكا?) كما سنرى في الفصل الرابع [4].

تتكو ? الإشارة الواحدة م? عناصر تسمى معالم الإشارة: شكل اليد، واتجاهها، ومكا? أدائها أمام الجسم، وحركتها عبر الزم?، ثم تعبيرات الوجه التي تحمل جـــزءا م? النحو نفسه (كالنفي والاستفهام). هذا التركيب يفسر لماذا تفشل الأنظمة التي تنظر إلى صورة واحدة ثابتة: إشارتا ? قد تبدآ ? بشكل اليد نفسه وتختلفا ? بالحركة فقط. كما يفسر لماذا اخترنا أ?@خذ نموذجنا تسلسلً م? 32 إطارا زمنيا لا إطارا واحدا، وأ? تتضم? خصائصه نقاطا مختارة م? الوجه إلى جانب اليدي? [6].

وتتميز لغة الإشارة العربية بتعدد لهجاتها بي? البلدا?، وبوجود طبقة مفردات صحية ويومية واسعة الاستعمال في المواقف الخدمية. وقد انعكس هذا على اختيارنا لمصدر البيانات: مجموعة KArSL سُجلت بمحدي? محترفي? وبإشارات موحدة نسبيا، وهذا يجعلها أساسا جيدا للتدريب، لكنه يعني أيضا أ? أداء النموذج على لهجة محليـــــة مختلفة سيبقى أقل م? أدائه المخبري، وهي نقطة عدنا إليها بصراحة عند مناقشة حدود النظام [4].

2-2

الرؤية الحاسوبية ونقاط المعلم

أمام أي نظام يريد فهم حركة الإنسا? م? الكاميرا طريقا?: إدخال بكسلت الصورة مباشرة إلى شبكة عصبية التفافية، أو استخراج تمثيل وسيط مدمج أولاً ثم التصنيف عليه. الطريق الأول أثقل حسابيا بكثير ويحتاج بيانات ضخمة، وحي? جرّبنا التفكير فيه لهاتف عادي وجدناه طريقا مسدودا عمليا. أما الطريق الثاني فيعتمد على مـــا يسمى نقاط المعلم (Landmarks): مجموعة إحداثيات ثلثية الأبعاد لنقاط مفصلية محددة في اليد والجسم والوجه، تخرج م? كل إطار فيديو. بهذا تتحول صورة حجمها مئات الكيلوبايتات إلى متجه م? 378 عددا فقط، ويصبح التصنيف م@لة أخـــف بمرتبات [7].

يبي? الشكل (1-2) مخطط نقاط معلم اليد الواحدة كما يستخرجها متعقّب اليد: 21 نقطة، أربع لكل إصبع (قاعدة ومفصل? وطرف) ونقطة للمعصم. م? هذه النقاط وحدها يمك? اشتقاق خصائص غنية: حالة كل إصبع (ممدودة أم مثنية) م? زوايا مفاصلها ومسافات أطرافها، والمسافات النسبية بي? الأصابع، واتجاه الكف. وعلى هذه الخصائص بالذات بُني وضع الإشارات السهلة في نظامنا بقواعد هندسية مباشرة دو? أي تدريب، وبُنيت أيضا شبكة الحروف التي@خذ 63 عددا هي إحداثيات نقاط يد واحدة [7].



رسم تخطيطي مبسّط: ثلاث نقاط ممثلة لكل إصبع من أصل أربع في النموذج الحقيقي

الشكل (1-2): مخطط مبسط لنقاط معلم اليد الواحدة وعددها 21 نقطة.

يظهر في المخطط توزيع النقاط على الأصابع الخمسة والمعصم، وهو توزيع موحّد في أدوات التتبع الحديثة يجعل الخصائص المستخرجة منه قابلة للنقل بي? التطبيقات ومجموعات البيانات. وتنبيه صغير يستحق الذكر: المخطط مبسط عمدا (ثلث نقاط ظاهرة لكل إصبع) لتسهيل القراءة، بينما يعمل النظام فعليا بالنقاط الإحدى والعشري? كاملة وبإحداثياتها الثلثية.

3-2مكتبة MediaPipe ونماذج التتبع

اخترنا لبناء طبقة التتبع مكتبة MediaPipe م? Google، وهي إطـــار مفتـــوح المصدر لبناء خطوط معالجة إدراكية تعمل على الجهاز نفسه، وقد صُممت أصلً لتشغيل نماذج تعلم آلي على الهواتف بسرعة لحظية [8]. يهمنا منها نموذجا?. الأول نموذج اليد (Hand Landmarker) الذي يكشف اليد ويخرج نقاطهـــا الإحـــدى والعشري?، وهو نموذج صغير الحجم (7.8 ميغابايتات) يعمل بمرحلتي?: كاشف كف يجد موضع اليد في الإطار، ثم نموذج معلم يخرج الإحداثيات الدقيقـــة، وقـــد بيّ?

منشوروه أنه يعمل لحظيا حتى على وحدات معالجة الرسوميات في الهواتف [7]. والثاني الحل الشامل (Holistic) الذي يجمع في ممر واحد نقاط الوضـــعية (33 نقطة) وشبكة الوجه (468 نقطة) ونقاط اليدي?، ليصل مجموع مـــا يخرجـــه إلى 533 نقطة لكل إطار، فاخترنا منه لمتجه الخصائص ما يلزم فقط كما @تي [9].

سبب تفضيلنا حلً جاهزا للتتبع على تدريب متعقّب خاص بنا عملي بحت: بياناتنا العربية تصف الإشارات لا هندسة اليد العامة، وتدريب متعقّب يد ينافس ما لـــدى Google يحتاج بيانات وحواسيب لا نملكها. بهذا التقسيم ركّزنا جهدنا البحثي حيث توجد مساهمتنا فعلً: في تصنيف الإشارات العربية وصياغة جملها، وتركنا مشكلة التتبع العامة لأداة ناضجة مجرّبة على ملييي? الأجهزة [8]، [9].

2-4 تصنيف التسلسلت الزمنية للشارات

إشارة الكلمة الواحدة حدث زمني: تبدأ حي? ترتفع اليدا? ع? وضع الراحة وتنتهي حي? تعودا? إليه. لذلك يتعامل نظامنا مع مقطع زمني طوله ثابت هو 32 إطارا، يُبنى لكل إطار فيه متجه خصائص م? 378 عددا، ثم يدخل التسلسل كاملً إلى شبكة التفافية زمنية (Temporal Convolutional Network) م? أربع كتل التفافية أحادية البعد بعرض 192 قناة، تتعلم الأنماط الحركية عبر الزم? وتخرج احتمالاً لكل إشارة في المفردات. اختيار الالتفافات الزمنية على الشبكات التكرارية كـــا? لسببي?: تدريب أسرع واستقرار أعلى على بيانات بهذا الحجم، وزم? استدلال قصير جدا على الجهاز بلغ نحو 1.5 ميلي ثانية للمقطع [6].

ويلخص الجدول (1-2) مكوّنات متجه الخصائص لكل إطار، وهو متجه يجب أ? يبقى متطابقا تماما بي? بيئة التدريب (Python) وبيئة التشـــغيل (JavaScript في المتصفح)؛ فاختلف ترتيب عنصر واحد يحوّل النموذج إلى مولّد تخمينات. لهـــذا السبب كتبنا ملفي الخصائص في البيئتي? توأمي? حرفيي?، واعتبرنا أي تعديل على أحدهما ملزما بتعديل الآخر في الالتزام نفسه.

الجدول (1-2): مكونات متجه الخصائص لكل إطار وعددها 378 عددا.

المكون	عدد النقاط	القيم لكل نقطة	المجموع
وضعية الجسم	33	4 (س، ص، ع، ظهور)	132
اليد اليمنى للمحدي	21	3 (س، ص، ع)	63
اليد اليسرى للمحدي	21	3 (س، ص، ع)	63
نقاط مختارة م? الوجه	40	3 (س، ص، ع)	120
المجموع			378

نلحظ أ? نقاط الوجه المختارة ليست الشبكة الكاملة؛ اخترنا 40 نقطة تحمل المعنى في الإشارة: محيط الشفتي?، والحاجبي?، والعيو?، وطرف الأنف والذق?. فتعبيرات الوجه جزء م? نحو لغة الإشارة، لك? إدخال 468 نقطة كاملة كا? سيضخّم المتجه ويعلّم النموذج ضجيجا لا معنى. وعند غياب أي مكوّ? في إطار ما (يد خارج الكادر مثلً) تُ مل خاناته بالأصفار، وهو اتفاق تعلّمه النموذج أثناء التدريب لأ? البيانـــات نفسها فيها أطر بيد واحدة.

وقبل التصنيف@تي خطوة التقطيع (Segmentation): كيف يعرف النظام أي? تبدأ الإشارة وأي? تنتهي في بث كاميرا مستمر؟ اعتمدنا قاعدة حركية بسيطة: الإشارة قائمة ما دامت اليدا? فوق خط راحة متعلم م? وضع الجسم، وتنتهي حي? تنخفضا? عنه وتستقرا?. هذه البساطة مقصودة؛ فالتقطيع المعقد قد يبتلع إشارات أو يقطعها، بينما قاعدة واضحة يستطيع المستخدم نفسه أ? يتحكّم بها فيضبط إيقاع أدائه عليها.

2-5مجموعات البيانات التدريبية

لا نموذج بل بيانات، والبيانات العربية هنا هي عنـــق الزجاجـــة الحقيقي. يلخص الجدول (2-2) المجموعتي? اللتي? اعتمدنا عليهما وخصائص كل منهما كمـــا هي موثقة في مصدريهما [4]، [5].

الجدول (2-2): مجموعتا البيانات المستعملتان في تدريب نماذج النظام.

الوجه	KArSL	ArASL2018
نوع المحتوى	مقاطع فيديو لإشارات كاملة	صور ثابتة لأشكال الحروف
الحجم الكلي	502 إشارة × 50 تكـــرارا × 3 مجدي?	54,049 صورة
عدد الفئات المستعملة	100 ثم 261 إشارة	32 حرفا
عدد المجدي?	3 محترفي?	40 مشاركا
أداة التسجيل	Kinect V2	كاميرات عادية
الترخيص	يُطلب الاستشـــهاد بالورقـــة، ولا ترخيص تجاريا معلنا	CC BY 4.0

أهم ما في هذا الجدول هو الرقم ثلثة: عدد مجدي KArSL. فجي? تدرّب نموذج على مجدٍّ واحد أو اثني? فإنه يحفظ أسلوبهما لا الإشارة ذاتها، وقد رأينا ذلك@عيننا في بدايات المشروع: نموذج درّبناه على 800 مقطع م? مجدٍّ واحد أعطى 65% فقط جي? جرّبناه على شخص آخر. لهذا قسنا كل نتائجنا بمنهجية استبعاد مجدٍّ كامل (Leave-One-Signer-Out): ندرّب على اثني? ونختبر على الثالث الذي لم يـــره النموذج قط، فتكو? النتيجة تقديرا أمينا لما سيحدث مع مستخدم جديد. أما مجموعة الحروف فل تحدُّد هوية المجدي لكل صورة، لذا قسمت عشوائيا، ونتيجة حروفنا (89.1%) قد تكو? متفائلة نسبيا لمستخدم جديد، وقد ذكرنا ذلك في موضـــعه دو? تجميل [5].

2-6 من الشارة إلى الكلم: الصياغة والنطق

آخر حلقات السلسلة حلقتا? لغويتا?. الأولى صياغة الجملة: خرجنا بقناعة مبكرة أ? نموذجا لغويا عاما سيكو? هنا عبئا لا عونا؛ فمفرداتنا معروفة ومغلقة (261 إشارة مصنفة)، وقواعد وصلها يمك? كتابتها صراحة بدقة أعلى وبل أي احتمال هلوسة، وبالمجا? ودو? اتصال. لذلك بنينا محرك قوالب: فئة «حاجة» مثل «صداع» تسبقها «عندي»، وفئة «فعل قصد» تليها «أريد الذهاب إلى» قبل المكـــا?، والحـــروف

المتتابعة تلتحم اسما واحدا. والثانية النطـــق: تحويـــل النص إلى كلم (Text-to-Speech) بجودة عربية مقبولة، وهي مشكلة مستقلة فا@تنا بصعوبتها، وسنروي في الفصلي? الثالث والرابع كيف جرّبنا صوتا مفتوح المصدر فرفضـــناه لجودتـــه، ثم استقرينا على مقاطع صوتية مولُّدة مسبقا تشغّل محليا، مع صوت الجهاز احتياطا أخيرا، وسنفصّل تجربة الصوت كاملة في الفصل الثالث.

الفصل الثالث:

الدوات والتقنيات المستخدمة

3-1 بيئة التطوير ولغة JavaScript وأداة Vite

بُنيت واجهة النظام بلغة JavaScript خالصة دو? إطار عمل (Framework) ثقيل، وكا? هذا قرارا واعيا لا قصورا: فحمل مكتبة واجهات كاملة فوق نماذج تعلم آلي يضخّم التحميل الأول على هواتف مستخدمينا واتصالاتهم، بينما واجهتنا شاشة واحدة تقريبا. استعملنا أداة Vite بوصفها خادم تطوير وبنّاء حزم، وهي أداة حديثة تجمع ملفات المشروع وتنتج نسخة إنتاج مضغوطة، وتدعم وحدات JavaScript الأصلية في المتصفح أثناء التطوير فيكو? التعديل والتجربة فوريي? [10]. واعتمدنا خـــط Alexandria العربي المتغير ليظهر النص العربي في الواجهة بخط واضح مقروء على الشاشات الصغيرة.

أما الوصول إلى الكاميرا فيتم عبر واجهة برمجة التطبيقات القياسية في المتصفحات (getUserMedia) التي تطلب إذ? المستخدم صراحة قبل تشغيل الكاميرا، ولا تعمل إلا عبر اتصال آم? (HTTPS)، وهو ما توفره منصة النشر كما س@تي [11]. هذه التفاصيل الصغيرة ليست هامشية: فإذ? الكاميرا هو لحظة الثقة الأولى بي? التطبيق ومستخدمه، وقد صمّمنا رسالة طلب الإذ? في الواجهة لتشرح بجملة واحدة أ? الفيديو يبقى على الجهاز.

3-2مكتبة MediaPipe Tasks Vision

طبقة التتبع في المتصفح هي حزمة MediaPipe Tasks Vision بالإصـــدار 1.0.1، ومنها نستعمل مهمتي?: متعقّب اليد (Hand Landmarker) لوضـــعي الإشارات السهلة والحروف، والحل الشامل (Holistic Landmarker) لوضـــع الكلمات الذي يحتاج اليدي? والوضعية والوجه معا [12]. تعمل هذه المهام بتقنيـــة WebAssembly مـــع تســـريع اختيـــاري بوحـــدة معالجـــة الرســـوميات،

وتُحمَّل نماذجها مرة واحدة ثم تخزّ? في ذاكرة التطبيق المحقتة لتعمل بعـــدها دو? اتصال [8].

اختيار نموذجي? مختلفي? للمهمة الواحدة تقريبا — تتبع اليد — لم يك? ترفا: فالحل الشامل أثقل بكثير، وتشغيله في وضع الإشارات السهلة الذي يكفيه كف واحدة كا? سيستنزف بطارية الهاتف بل مقابل. القاعدة التي اتبعناها في المشروع كله: لا تشغّل نموذجا أثقل مما تتطلبه المهمة الحالية، وأوقف كل ما لا يلزم فور انتقال المستخدم إلى وضع آخر.

3-3صيغة ONNX ومشغل ONNX Runtime Web

صدّرنا نموذج الكلمات المدرب إلى صيغة ONNX، وهي صيغة تبادل مفتوحـــة لتمثيل نماذج التعلم الآلي مستقلة ع? إطار التدريب الذي أنتجها، بحيث يمك? تشغيل النموذج نفسه في بيئات ولغات مختلفة دو ? إعادة تدريب [13]. يشغّل النموذج داخل المتصفح مشغّل ONNX Runtime Web بالإصدار 1.29.0، وهو النســـخة المخصصة للويب م? مشغّل ONNX Runtime، وينفّذ الاستدلال على وحـــدة المعالجة المركزية عبر WebAssembly [14].

قياسنا العملي أكد صحة هذا الاختيار: ملف النموذج نحو 3 ميغابايتات لفئات بلغت 261 إشارة وبنحو 700 ألف معامل، وزم? الاستدلال للمقطع الواحد بحدود 1.5 ميلي ثانية على وحدة معالجة مركزية عادية، أي أ? التصنيف ليس هو عنق الزجاجة إطلقا؛ الزم? الذي يشعر به المستخدم هو زم? الكاميرا واستخراج النقاط. واستثناء واحد جدير بالذكر: شبكة الحروف لم نصدّرها إلى ONNX أصلً، بل كتبنا تنفيذها الاستدلالي بلغة JavaScript مباشرة م? أوزانها المخزنة بصـــيغة ثنائيـــة (353 كيلوبايتا)، لأ? شبكة صغيرة م? ثلث طبقات لا تحتاج مشغّلً كاملً، فوفّرنا حجم تحميله على وضع الحروف كله [14].

3-4

بيئة التدريب: Python ومكتبة PyTorch

على الطرف الآخر م? خط العمل تقف بيئة التدريب بلغة Python ومكتبة PyTorch، وهي مكتبة تعلم عميق واسعة الانتشار في البحث العلمي لمرونتها وسهولة تصحيح أخطائها [15]. فيها كتبنا أدوات استخراج النقاط م? مقاطع KArSL المصورة، وبناء متجهات الخصائص، وتدريب شبكتي النظام، وتصدير النموذج إلى ONNX. واخترنا PyTorch تحديدا لأ? تصدير ONNX منها مباشر، ولأ? كود التدريب والاستدلال يبقى قريبا م? صيغة البحث الأكاديمي فيسهل على أي باحث إعادة إنتاج نتائجنا [15].

وفي بيئة التدريب نفسها وُلدت أداتا? إضافيتا? صارتا جزءا م? المنتج النهائي لا م? المختبر فقط: أداة توليد رسوم القاموس المتحركة م? أنظف مقاطع البيانات، وأداة توليد المقاطع الصوتية للمفردات. يلخص الجدول (1-3) الأدوات والتقنيات المستعملة في المشروع وإصداراتها كما هي مثبتة في ملفات المشروع نفسه.

الجدول (1-3) : الدوات والتقنيات المستعملة في بناء النظام وتشغيله.

الداة أو التقنية	الصدار	دورها في النظام
Vite	7.3.6	بناء حزمة الويب وخادم التطوير
MediaPipe Tasks Vision	1.0.1	تتبع اليد والحل الشامل داخل المتصفح
ONNX Runtime Web	1.29.0	تشغيل نموذج الكلمات على الجهاز
خط Alexandria المتغير	5.3.0	خط الواجهة العربية
Wrangler	4.149.0	نشر التطبيق إلى منصة Cloudflare Pages
PyTorch	بيئة التدريب	تدريب الشبكتي? وتصدير ONNX
NumPy	بيئة التدريب	معالجة متجهات الخصائص والمقاطع

3-5

توليد الصوت: من Piper إلى المقاطع الجاهزة

جرّبنا أولاً توليد النطق محليا@داة Piper المفتوحة المصدر، وهي محرك تحويل نص إلى كلم عصبي سريع يعمل على الجهاز، واخترنا صوته العربي المتـــاح (ar_JO-kareem-medium) [16]. لكـ? التجربة العملية حسمت الأمر بسرعة: جودة الصوت العربي الناتج لم تكـ? مقبولة لمستخدم سيسمع جُمله في مكا? عام؛ نطق متقطع ولكنة غريبة تجعل الرسالة الصحية الجادة تبدو غير موثوقة. فرفضناه رغم جاذبية كونه مجانيا ومحليا بالكامل، لأ? جودة الصوت في نظامنا ليست زخرفة بل هي الرسالة نفسها.

البديل الذي استقر عليه المشروع: توليد مقاطع صوتية جاهزة لكل مفردة وقالب جملة مرة واحدة أثناء الإنتاج، باستعمال خدمة ElevenLabs بصوتها المسمى George مـ? نموذجها المتعدد اللغات، بعد مقارنة أربعة أصوات بتحويل نطقها آليا إلى نص (Whisper) وقياس أخطائها، فكا? هذا الصوت وصوت آخر بل أخطاء واخترناه لطبقته الأوضح [17]. تُشغّل هذه المقاطع محليا أثناء الاستعمال فل يحتاج النطق اتصالاً، وحدود الخطة المجانية للخدمة (10,000 حرف شهريا) استُهلك منها نحو النصف في توليد مفرداتنا، وهو قيد سجّلناه لأنه يحدّ مـ? سرعة توسيع المفـــردات مستقبلً. وعند غياب مقطع جاهز لكلمة ما، يتراجع النظام إلى صوت الجهاز العربي عبر واجهة النطق في المتصفح (SpeechSynthesis) بوصفه حلً أخيرا لا أساسيا [18].

3-6النشر: Cloudflare Pages وتطبيق الويب التقدمي

يُنشَر التطبيق على منصة Cloudflare Pages بوصفه موقعا ثابتا بالكامل: لا خادم خلفي ولا قاعدة بيانات ولا واجهات برمجية على الشبكة، وهذا انسجام مقصود مع

قرار «كل شيء على الجهاز». النشر سطر واحد عبر أداة Wrangler يرفع مجلد البناء النهائي، والموقع متاح على نطاق sign-to-speech.pages.dev [19].

وفوق ذلك بنينا التطبيق تطبيق ويب تقـدميا (PWA): عامـل خدمـة (Service Worker) يخزّ? عند أول زيارة ملفات التطبيق والنماذج والأصوات، فيفتح التطبيق بعدها فورا ويعمل دو? اتصال، وقابل للتثبيت على الشاشة الرئيسة@ي تطـبيق أصلي [20]. عامل الخدمة هذا كا? أيضا مصدر أصعب خ@ واجهناه في النشر كله: نسخة مبكرة منه جعلت الموقع لا يفتح دو? اتصال كما وعدنا، ولم يُكتشف السبب إلا بمراجعة طريقة تخزي? أصول النماذج الكبيرة،@صلحناه وأعدنا النشر. الدرس الذي خرجنا به وسجّلناه: ميزة «يعمل دو? اتصال» لا تُختبر م? بيئة التطوير، بل م? هاتف حقيقي في وضع الطيرا?.

الفصل الرابع:

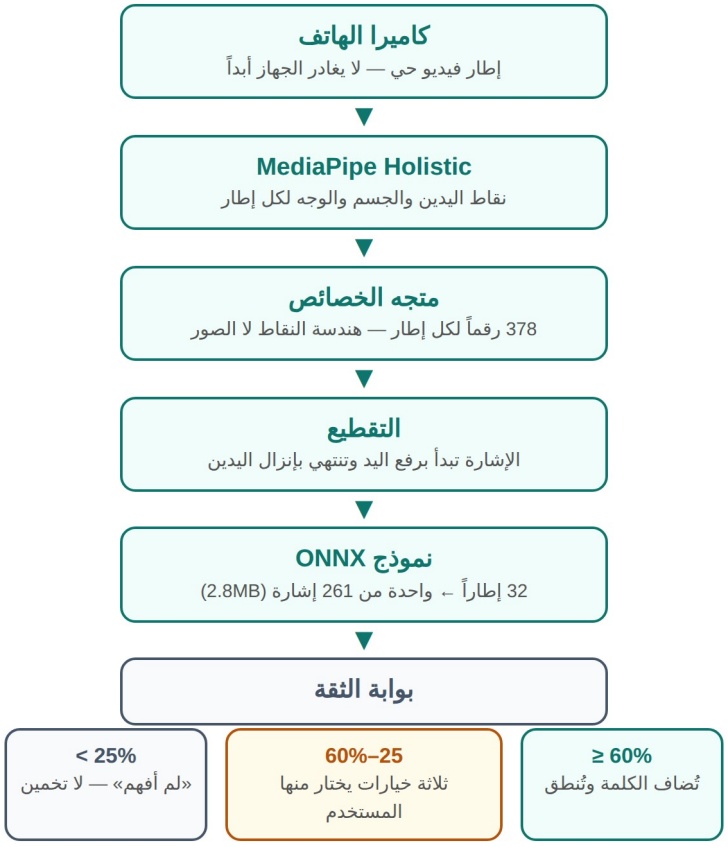
التطبيق العملي

4-1 مقدمة

في هذا الفصل ننزل م? النظرية إلى الورشة: كيف بُني النظام فعليا، وما القرارات التي اتخذناها عند كل مفترق، وما الأرقام التي حصلنا عليها حي? اختبرناه بمنهجية صارمة. ترتيب الفصل يتبع ترتيب سير البيانات في النظام نفسه: م? المعمارية العامة، إلى أوضاع التعرف الثلثة، إلى خط التدريب، إلى صياغة الجملة، إلى القاموس، ثم النتائج والحدود بصراحة كاملة.

4-2معمارية النظام وخط المعالجة

يبي? الشكل (4-1) خط المعالجة الكامل لوضع الكلمات، وهو أعقد أوضاع النظام وقلبه. تبدأ الرحلة م? كاميرا الهاتف، فيستخرج الحل الشامل م? كل إطار نقاط اليدي? والوضعية والوجه، وتُبنى منها متجهات الخصائص ذات الأعداد 378، ثم تراقب وحدة التقطيع حركة اليدي? بالنسبة إلى خط الراحة فتقتطع تسلسل الإشارة وتمدّده أو تضغطه إلى 32 إطارا، ويدخل التسلسل إلى نموذج ONNX فيخرج بتوزيع احتمالات على المفردات، فتطبّق بوابة الثقة قرارها: إضافة ونطق مباشر عند ثقة 60%@كثر، أو عرض ثلثة خيارات للمستخدم بي? 25% و60%، أو رسالة «لم أفهم» دو? ذلك. وفي تطور لاحق م? تجربة الاستعمال جعلنا الوضع يعرض أقرب تخمي? مباشرة إذا تجاوز 15% مع تنبيه المستخدم أ? بوسعه حذف الكلمة بزر الحذف إ? أخ@ت، لأ? إيقاف الحوار لعرض خيارات في كل مرة كا? يبطئ المحادثة الحقيقية.



الشكل (1-4) : خط معالجة وضع الكلمات من الكاميرا حتى إضافة الشارة إلى الجملة.

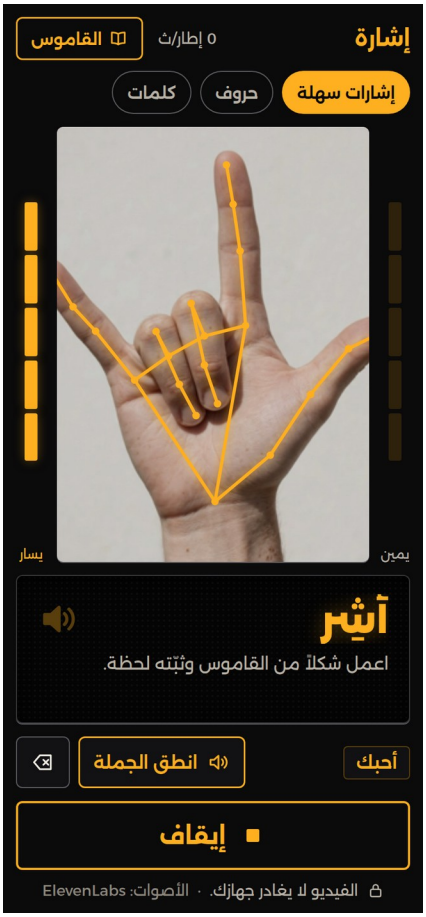
نلحظ في المخطط أ? كل المراحل تجري داخل صفحة المتصفح الواحدة، بل أي طلب شبكي بعد التحميل الأول. هذا التصميم له ثم? معروف دفعناه بوعي: حجم تحميل أول أكبر (النماذج والأصوات)، مقابل صفر كلفة تشغيل خادم، وخصوصية مطلقة، وعمل بل اتصال. وللمقارنة، تعمل الأوضاع الأخف (الإشارات السهلة والحروف) بخطوط أقصر م? الخط نفسه: متعقّب اليد وحده بدل الحل الشامل، وقواعد أو شبكة صغيرة بدل نموذج ONNX.

4-3 وضع الشارات السهلة

وضع الإشارات السهلة هو الباب الذي يدخل منه كل مستخدم جديد، وقد وُلد م? تجربة ميدانية: حي? جرّب صاحب المشروع النموذج الأول بنفسه وجد أ? إشارات KArSL المتحركة صعبة الأداء والتعرف على المبتدئ،@ضفنا وضعي? بيد واحدة،

وجعلنا الأسهل هو الافتراضي. يعمل هذا الوضع بتسع إشارات ثابتة الشكل يقرها النظام بقواعد هندسية على حالة كل إصبع (ممدودة، مثنية، بينهما) مستخرجة م? نقاط اليد الإحدى والعشري?، دو ? أي نموذج مدرّب، ولذلك يعمل مع أي شخص م? أول مرة مهما كانت يده أو لهجته.

ولكل إشارة قاعدة تثبيت زمنية: لا تُسجَّل الإشارة إلا إذا بقي شكل اليد ثابتا 550 ميلي ثانية يمتلئ خللها شريط تقدم أمام المستخدم، ثم لا تتكرر الإشارة نفسها حتى يتغير الشكل أو تنخفض اليد، حتى لا تمتلئ الجملة بتكرارات غير مقصودة أثناء الأداء الطبيعي. وقيست القواعد لاحقا على صور مشابهة م? مجموعة الحروف فجاءت نسب تعرّف الكف المفتوحة (مرحبا) بي? 85% و92%، بينما كانت أشكال أدقّ مثل إبهام «نعم» أصعب على القياس بنسبة 58%، وهي أرقام عاملناها كمؤشر جودة للقواعد لا كدقة نموذج. ويبي? الشكل (2–4) لقطة م? اختبارنا الحي لهذا الوضع بكاميرا وهمية، إذ تعرّف النظام على شكل «أحبك» وأضافه إلى الجملة.



الشكل (2–4) : تعرف النظام على إشارة «أحبك» في وضع الشارات السهلة أثناء اختبار حي بكاميرا وهمية.

تظهر اللقطة لوحة الجملة وقد سُجلت فيها كلمة «أحبك» مع رسالة@كيد الإضافة، وشريط الأوضاع الثلثة أعلى الشاشة. وقد سمحت لنا طريقة الاختبار هذه — تغذية الكاميرا الوهمية بصور مولّدة لإشارات معروفة — بالتحقق م? سلسلة العمل كاملة (كاميرا، تتبع، قواعد، جملة) دو ? الحاجة إلى محدٍّ بشري في كل مرة، على أننا نحكد أ? الصورة المولّدة يد نموذجية واضحة، فهي تثبت سلمة الخط لا دقته على أيدٍ حقيقية متنوعة، وهي نقطة ناقشناها في حدود النظام.

أما حساب حالة الإصبع نفسه فيتم بدالتي? صغيرتي? كتبناهما بعناية، لأ? تجربة الأرقام علمتنا أ? زوايا مفاصل الإبهام تحديدا تخدع حي? يشير إلى الكاميرا؛ فقِيس انفصال الإبهام بمسافة طرفه ع? قاعدة البنصر بدل زواياه. فيما@تي مقتطف فعلي م? شيفرة الوضع يوضح الفكرة:

```
function fingerState(f, [mcp, pip, dip, tip]) {
 const [m, p, d, t] = [pt(f, mcp), pt(f, pip), pt(f, dip), pt(f, tip)];
 const w = pt(f, 0);
 const straight = cos(sub(p, m), sub(t, p));
 const reach = dist(t, w) / (dist(p, w) || 1);
 if (straight > 0.75 && reach > 1.25) return 1;
 if (straight < 0.35 || reach < 1.05) return 0;
 return 0.5;
}
```

يجمع هذا المقتطف بي؟ مقياسي؟: استقامة الإصبع م؟ جيب تمام زاوية مفاصلها، وامتدادها م؟ نسبة بعد طرفها ع؟ المعصم إلى بعد مفصلها الأول عنه؛ فإذا اجتمع الامتداد والاستقامة عُدّت ممدودة، وإلا فمثنية أو وسطى. وفوق القواعد أضفنا طبقة تخصيص: يستطيع المستخدم إعادة تعيي؟ معنى أي إشارة سهلة، وتعليم التطـــبيق إشارة جديدة خاصة به بتسجيل 2.5 ثانية بعد عدٌّ تنازلي، تحفظ نقاطها فقـــط في تخزي؟ المتصفح المحلي، ويُقار ؟ الأداء الجديد@قرب جار ضم؟ مسافة قصـــوى مضبوطة جعلت نسبة القبول الخاطئ بي؟ الأشخاص بحدود 3%. وأخيرا، تُكشف حركة التلويح (لإشارة «مع السلمة») م؟ تذبذب الكف المفتوحة زمنيا لا م؟ شكل ثابت.

4-4وضع الحروف

حي؟ لا توجد إشارة لكلمة ما — اسم شخص، اسم مدينة، دواء جديد — يبقى الحل الذي يعرفه كل مستخدمي لغة الإشارة: التهجئة بالأصابع. يدعم وضع الحـــروف أشكال الأبجدية العربية الاثني؟ والثلثي؟، ويتعرف عليها بشبكة عصبية صغيرة م؟ ثلث طبقات (63 مدخلً هي إحداثيات يد واحدة، ثم طبقتا؟ مخفيتا؟ بعرض 256، ثم 32 مخرجا)، حجم أوزانها 353 كيلوبايتا فقط وتنفَّذ بلغة JavaScript مباشرة بل مشغّل نماذج. درّبناها على صور مجموعة ArASL2018 بعد تحويلها إلى نقاط يد، فبلغت دقتها 89.1% على 10,208 صورة محتجزة للختبار، وكـــانت أضـــعف حروفها القاف (77%) والذال (76%) والزاي (80%)، وهي حـــروف تتشـــابه أشكالها فعلً حتى على العي؟ البشرية [5].

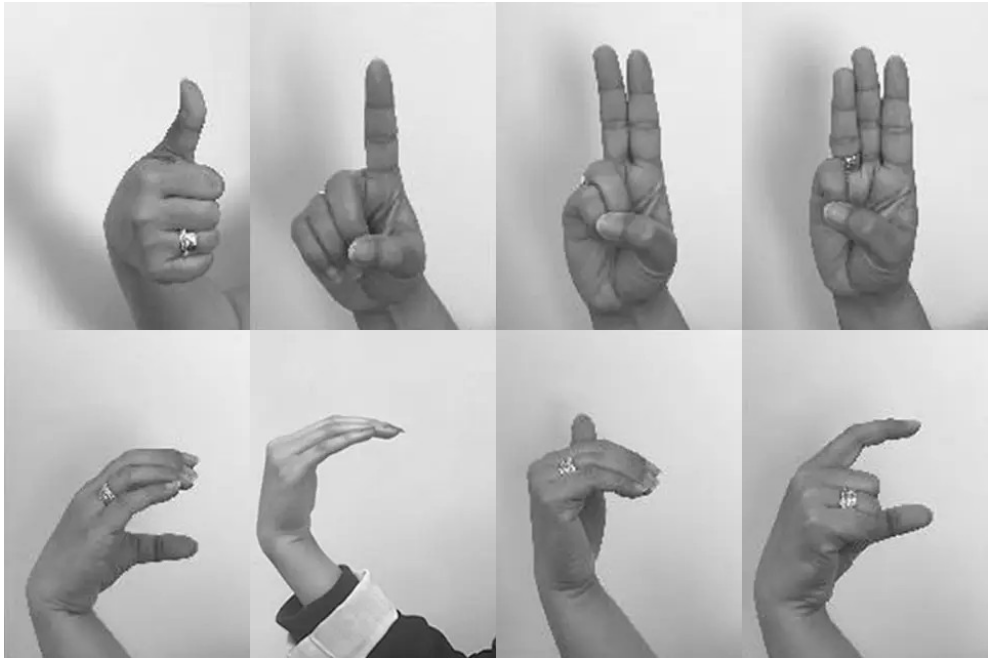
زم? تثبيت الحرف أطول قليلً م? الإشارة السهلة (650 ميلي ثانية) لأ? التهجئة تتطلب إيقاعا أبط@، وتتجمع الحروف المتتابعة تلقائيا في كلمة واحدة داخل الجملة. ويبي? الشكل (3-4) شاشة دليل الحروف داخل التطبيق، وقد استبدلنا فيها رسوما هيكلية جرّبناها أولاً ولم يفهمها المستخدمو? بصور حقيقية م? مجموعة البيانـــات نفسها لكل حرف.



الشكل (3-4) : دليل حروف البجدية الشارية داخل التطبيق بصور حقيقية وزر تجربة لكل حرف.

في هذه الشاشة يرى المستخدم صورة كل حرف كما يحديه شـــخص حقيقي، وزر «جرّب» ينقله إلى الكاميرا ليحدي الشكل بنفسه ويتلقى تقييما فوريا. وقرار استبدال الرسوم الهيكلية بالصور الحقيقية لم يك? جماليا: فالرسم الهيكلي يريك أي? النقاط، لكنه لا يريك كيف تبدو اليد فعلً حي? تلتف الأصابع حول بعضها، وقد لاحظنـــا

ارتباك المستخدمي؟ معه في أول تجربة فغيّرناه دو ؟ تردد. ويبي؟ الشكل (4-4) عينة م؟ هذه الصور المرجعية كما هي في ملفات المشروع.



الشكل (4-4): صور مرجعية من مجموعة ArASL2018 لول ثمانية حروف في دليل التطبيق.

تُظهر العينة الحروف م؟ الألف إلى الدال بصورها الأصلية م؟ المجموعة، وهي الصور نفسها المعروضة داخل التطبيق بعد تحويلها إلى صيغة ويب خفيفة. ونشير إلى أ؟ حقوق هذه الصور تعود لمجموعة ArASL2018 المرخّصة بالمشاع الإبداعي مع وجوب الاستشهاد، وقد التزمنا بذلك داخل التطبيق نفسه في صفحة القاموس وفي مراجع هذا الكتاب [5].

4-5وضع الكلمات وخط التدريب

وضع الكلمات هو ذروة النظام وسبب وجوده: التعرف على إشارات متحركة كاملة م؟ مفردات واسعة. يبي؟ الشكل (4-5) خط التدريب الذي بنيناه لهذا الوضع م؟ الطرف إلى الطرف: مقاطع KArSL الخام تستخرج منها النقاط@داة استخراج، فتُبنى متجهات الخصائص، فتُدرّب الشبكة في PyTorch، فتُصدّر إلى ONNX،

فتُنسخ إلى مجلد التطبيق العام مع قاموس المفردات، فيعمل النموذج في المتصفح كما لو كا? جزءا م? الصفحة.

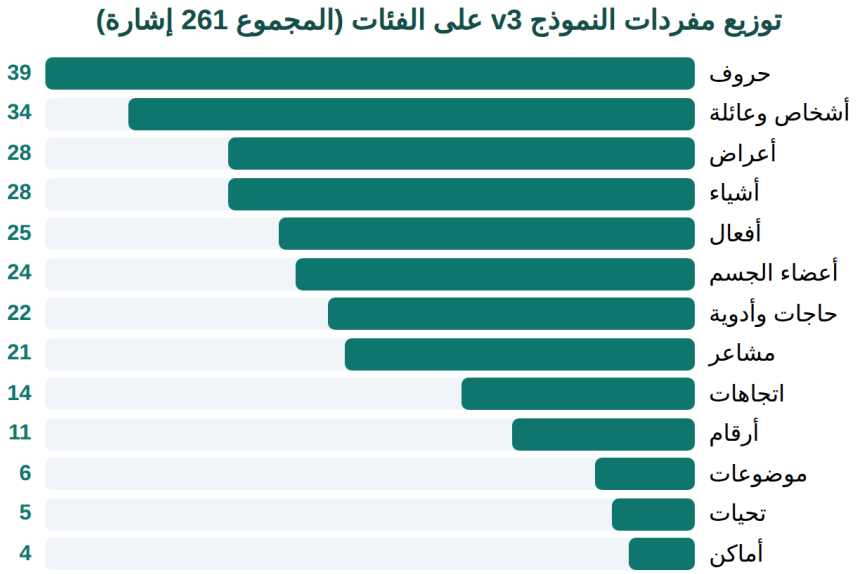


الشكل (5–4): خط تدريب نموذج الكلمات من مقاطع البيانات حتى النشر داخل التطبيق.

القاعدة الذهبية في هذا الخط هي التطابق التام بي? خصائص التدريب وخصائص التشغيل: ملف الخصائص في Python وتوأمه في JavaScript ينتجا? المتجه نفسه بالترتيب نفسه لأي إطار، وأي تعديل على أحدهما دو? الآخر يفسد النموذج بصمت. لهذا كتبنا في رأس الملفي? تنبيها صريحا بهذا المعنى، وجعلنا أي تغيير عليهما يمر بمراجعة مزدوجة. والنموذج نفسه شبكة التفافية زمنية م? أربع كتـــل، بعـــرض 192 قناة، دخلها تسلسل 32 إطارا × 378 عددا، وخرجها توزيـــع على فئـــات المفردات.

مرّ النموذج بمرحلتي? موثقتي?. في الأولى (الإصدار الثاني م? التطبيق) درّبناه على 100 إشارة صحية م? KArSL بمقاطع بلغت 4,800 مقطع م? المحدي? الثلثة، بعد أ? واجهنا قيدا عمليا يستحق التسجيل: حصّة التحميل اليومية م? خدمة التخزي? التي استضافت المقاطع الخام (4–6 غيغابايتات لكل محدٍّ) أوقفتنا، فجمعنا البيانات م? مصدري?: مجموعة الاختبار الرسمية م? الفيديو الخام للمـــحدي? الثلثـــة (8 تكرارات لكل إشارة)، ومجموعة التدريب بنسختها الإطارية 256×256 للمحدي? الأول والثاني (حتى 12 تكرارا). ولاحظنا فرقا مهما بي? المصدري?: اليد اليسرى تُكشف في 46% فقط م? أطر النسخة الإطارية مقابل 74% في الفيديو الخام، كما احتجنا حشو الأطر إلى نسبة عرض 16:9 لتطابق هندسة الكاميرا. وفي المرحلة الثانية وسّعنا المفردات إلى 261 إشارة بمقاطع بلغت 13,286 مقطعا،@صـــبح

النموذج الحالي (الإصدار الثالث) هو المنشور على الموقع. ويبي؟ الشكل (6–4) توزع المفردات الحالية على فئاتها الدلالية.



الشكل (6–4): توزع مفردات النموذج الحالي وعددها 261 إشارة على فئاتها.

تُظهر البيانات أ? المفردات ليست صحية فقط بعد التوسعة: فئات للشخاص (34) والأشياء (28) والأعراض (28) والأفعال (25) والجسم (24) والحاجـــات (22) والمشاعر (21)، إلى جانب الحروف (39) بوصفها فئات نموذج أيضا. هذا الاتساع كا? مقصودا لنقل التطبيق م? «أداة عيادة» إلى أداة محادثة يومية، لكنه جاء بثم? في الدقة سنعرضه في النتائج. ويقار? الجدول (1–4) بي? الأوضاع الثلثة م? حيث الأسلوب والحاجة.

الجدول (1–4) : مقارنة بين أوضاع التعرف الثلثة في النظام.

الوجه	الشارات السهلة	الحروف	الكلمات
الأسلوب	قواعد على شكل اليد	شـــبكة صـــغيرة 63→ 32→256→256	شبكة التفافيـــة زمنيـــة ONNX
المتعقّب	Hand Landmarker	Hand Landmarker	الحل الشامل Holistic
المفردات	9 إشارات قابلة للتعليم	32 حرفا	261 إشارة
زم? التثبيت	550 ميلي ثانية	650 ميلي ثانية	تقطيع بحركة اليدي?

الوجه	الشارات السهلة	الحروف	الكلمات
التدريب اللزم	لا شيء	ArASL2018	KArSL
أنسب استعمال	عبـــارات فوريـــة لأي شخص	تهجئة الأسماء والجديد	جمل م? مفردات واسعة

6-4صياغة الجملة والنطق

حي? تتجمع الإشارات في لوحة الجملة@تي دور محرك القوالب، ويبي? الشـــكل (4-7) مثاله الأشهر م? اختباراتنا: إشارتا «صداع» و«دواء» تتحولا? إلى جملة منطوقة كاملة «عندي صداع، أحتاج دواء.».



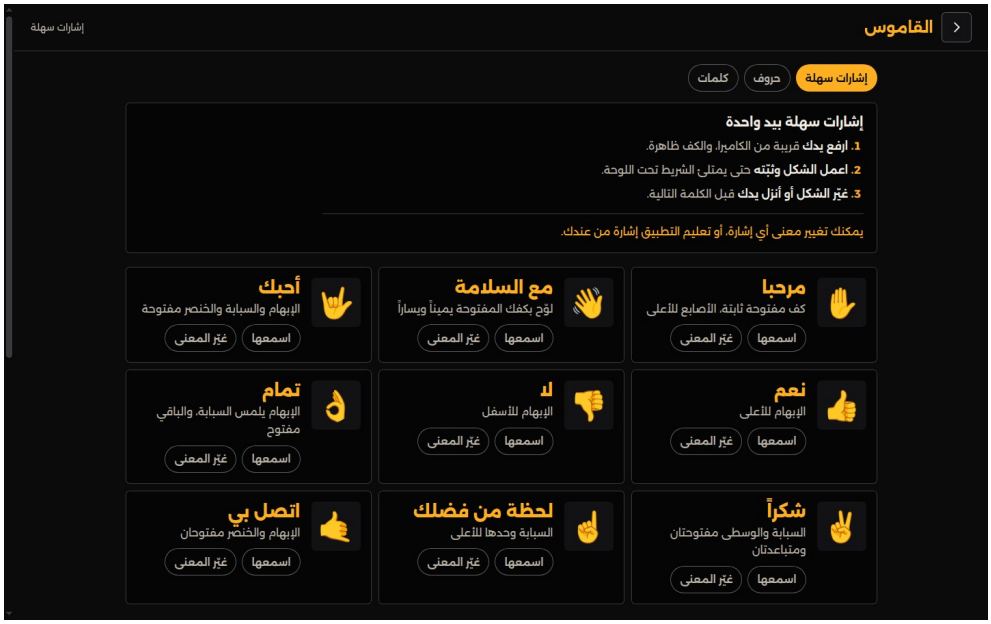
الشكل (4-7): مثال صياغة جملة من إشارتين بمحرك القوالب.

نلحظ أ? أيا م? كلمات الجملة النهائية («عندي»، «أحتاج») لم تُحدَّ بإشارة مستقلة؛ إنها وليدة فئتي الإشارتي?: «صداع» م? فئة الأعراض فتُفتتح بـــ«عندي»، و«دواء» م? فئة الحاجات فتُبنى بـــ«أحتاج». وقواعد المحرك مكتوبة صراحة في قـــاموس المفردات: لكل فئة قالبها، وللفعال صيغها («أريد الذهاب إلى» قبـــل الأمـــاك?)، وللعطف قاعدة وصل عربية سليمة («صداع وحمى وإسهال»)، وللحروف المتتابعة التحام في اسم واحد. اخترنا هذا الطريق واعيا بدل النماذج اللغوية العامة: فمـــع مفردات مغلقة معروفة تكو? القوالب أدقّ، ولا تهلوس أبدا، وتعمل بل خادم وبل كلفة. حدوده معروفة أيضا وسجلناها: لا يستطيع صياغة تراكيب خارج قوالبـــه، وسيحتاج توسعة يدوية مع كل فئة دلالية جديدة.

أما النطق فيعمل بسلّم أولويات: مقطع صوتي مولَّد مسبقا للكلمة إ? وجد، فصوت الجهاز العربي، فصوت Piper المحلي أخيرا. وقد اكتشفنا أثناء توليد المقاطع أ? الأرقام والحروف يجب أ? تُولّد@سمائها متبوعة بنقطة حتى لا تُقتطع نهاياتها صوتيا، وأ? فحص المقاطع آليا بإعادة تحويلها إلى نص كشف ضعف أسماء حروف بعينها (ق، خ، ظ، ث) لأ? الصوت المختار إنكليزي الأصل مهما أجاد العربية العامة، فقبلنا هذا القيد ووثّقناه بدل إخفائه.

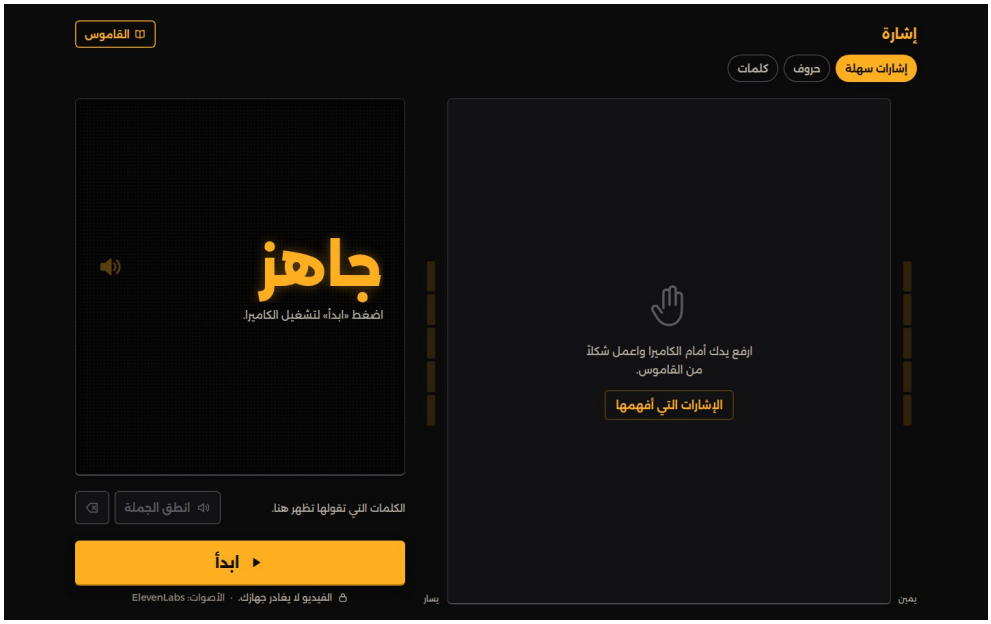
7-4القاموس والتدريب الذاتي

التطبيق الذي يطلب م? مستخدمه أداء إشارات لا يعرفها يكو ? قد حل نصف المشكلة فقط. لذلك بنينا داخل التطبيق قاموسا يعرض مفردات النظام كلها: لكل إشارة رسم متحرك مولّد م? أنظف مقطع حقيقي لها في البيانات (أداة خاصة استخرجت نقاط ذلك المقطع وحولتها إلى رسم هيكلي متحرك)، مع بحث نصي وتصفح بالفئات، وصفحة لكل إشارة فيها الرسم بحجم كبير وتشغيل بطيء وجملة مثال وزر لسماعها وزر «جرّبها بالكاميرا». ويبي? الشكل (8-4) شاشة القاموس كما تظهر على الحاسوب.



الشكل (8-4) : شاشة قاموس الشارات مع رسوم الحروف المتحركة وفئات المفردات.

تُظهر الشاشة شبكة رسوم الحروف المتحركة وفئات القاموس أسفلها، وهي البوابة التي يتعلم منها المستخدم مفردات النظام قبل أ? يحتاجها في موقف حقيقي. ووضع التدريب المرتبط بالقاموس يقيّم أداء المستخدم للشارة المختارة ويجيبه بـــ«أحسنت» أو «قريب» أو «حاول مرة أخرى» دو ? أ? يضيف شيئا إلى جملته، فهو مساحة آمنة للخ@. ويبي? الشكل (9–4) الواجهة الرئيسة للنظام على شاشة الحاســـوب حيث تُستعمل الكاميرا المدمجة أو الخارجية للتجربة والتدريب المطوّل.



الشكل (9–4): الواجهة الرئيسة للنظام على الحاسوب مع لوحة الجملة وأزرار التحكم.

على الحاسوب تتسع لوحة الجملة وتظهر عناصر التحكم بوضوح أكبر، وهذا الشكل هو الذي استعملناه في معظم اختبارات التطوير والكاميرا الوهمية، بينما تبقى تجربة الهاتف هي الأساس في التصميم لأنها بيئة الاستعمال الحقيقية لمستخدمنا.

8–4الاختبار والنتائج

قيّمنا النماذج بمنهجية استبعاد محدٍّ كامل: في كل جولة ندرّب على محديي? ونختبر على الثالث، ونحسب الدقة عند أول تخمي? (Top-1) وعند أول خمسة تخمينات (Top-5). يعرض الجدول (2–4) نتائج النموذج الأول ذي المئة إشارة.

الجدول (4-2): نتائج نموذج المئة إشارة باستبعاد مؤ كامل.

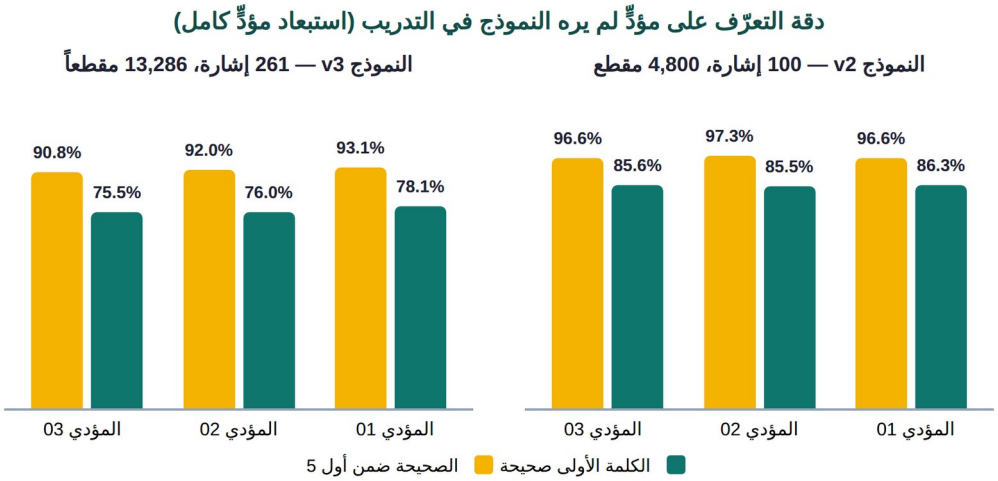
المؤدي المحتجز	مقاطع التدريب	مقاطع الاختبار	الدقة الولى %	ضمن أول خمسة %
الأول	2,800	2,000	86.3	96.6
الثاني	2,800	2,000	85.5	97.3
الثالث	2,800	2,000	85.6	96.6
المتوسط			85.8	96.8

حقق هذا النموذج هدف المرحلة الأولى الذي وضعناه لأنفسنا (85% على الأقل على محدٍّ جديد)، فانتقلنا إلى توسيع المفردات. ويعرض الجدول (4-3) نتائج النموذج الحالي ذي 261 إشارة بالمنهجية نفسها.

الجدول (4-3): نتائج نموذج 261 إشارة باستبعاد مؤ كامل.

المؤدي المحتجز	مقاطع التدريب	مقاطع الاختبار	الدقة الولى %	ضمن أول خمسة %
الأول	8,734	4,552	78.1	93.1
الثاني	8,726	4,560	76.0	92.0
الثالث	9,112	4,174	75.5	90.8
المتوسط			76.5	92.0

انخفضت الدقة الأولى م? 85.8% إلى 76.5% حي? كبرت المفردات 2.6 مرة، وهو انخفاض متوقع ومقبول عندنا لسببي?: عدد المقاطع لكل إشارة جديدة كا? أقل (فئات اليومية جاءت م? النسخة الإطارية بثمانية تكرارات فقط)، والإشارات المتشابهة حركيا تتكاثر مع اتساع القاموس. ويبي? الشكل (4-10) هذه المقارنة بصريا عبر المحدي? الثلثة.



الشكل (10–4): مقارنة دقة النموذجين باستبعاد مؤدٍّ كامل عبر المؤدين الثلثة.

نلحظ م? الرسم أ? الفجوة بي? النموذجي? ثابتة تقريبا عبر المحدي?، وأ? دقة «ضم? أول خمسة» بقيت فوق 90% في النموذج الأوسع، وهذا مهم عمليا لأ? نظامنا حي? يتردد يعرض خيارات؛ فوقوع الإشارة الصحيحة بي? أول خمسة تخمينات بنسبة 92% يعني أ? التصحيح م? المستخدم سريع حي? يلزم. أما التحقق الطرفي (End-to-End) فقد أجريناه على خط الأنابيب كاملً بتشغيل مقاطع حقيقية إطارا إطارا، فتعرف النظام على «قلب» و«صداع» و«دواء» و@كل» بنسب ثقة تجاوزت 91%، ثم أجرينا اختبار الكاميرا الوهمية على الموقع المنشور نفسه بمقطع «صداع» الحقيقي@ضيف إلى الجملة، وبمقاطع مرآتية متتابعة تعرف على «صداع، دواء، صداع، دواء» أربع مرات م? أربع، ونطق بعدها الجملة «عندي صداع، أحتاج دواء.» كاملة.

وأخيرا أعدنا اختبار النظام@نفسنا م? خارج بيئة التطوير الأصلية، ببناء نسخة الموقع محليا وتغذية كاميرتها الوهمية بصور مولّدة لإشارات سهلة معروفة، فجاءت النتائج كما في الجدول (4–4).

الجدول (4-4) : نتائج اختبار حي مستقل لوضع الشارات السهلة بكاميرا وهمية وصور مولدة.

الصورة المغذاة	شكل اليد	النتيجة على لوحة الجملة
إبهام وسبابة وخنصر ممدودة	شكل «أحبك» العالمي	أُضيفت «أحبك» مع رسالة@كيد
إبهام يلمس السبابة وباقي الأصابع مفتوحة	حلقة «تمام»	أُضيفت «تمام»
إبهام عند الأذ? وخنصر عند الفم	شكل «اتصل بي»	أُضيفت «اتصل بي»

ويبي? الشكل (4-11) لقطتي? م? هذا الاختبار جنبا إلى جنب.



الشكل (4-11): لقطتا اختبار حي: تعرف النظام على إشارتي «تمام» و«اتصل بي» من صور مولدة عبر كاميرا وهمية.

تُظهر اللقطتا? لوحة الجملة في كلتا الحالتي? وقد سجلت الكلمة الصحيحة، وم? المهم أ? نقرأ هذا الاختبار في حدوده: إنه يثبت سلمة سلسلة العمل كاملة م? الكاميرا إلى الجملة في الأوضاع الثلثة (فقد تحققنا في الجلسة نفسها م? وضع الحروف والوضع الافتراضي)، لكنه لا يقيس الدقة على أيدٍ بشرية حقيقية متنوعة، فالصور المولّدة

نظيفة الإضاءة والخلفية. قياس الدقة الحقيقي هو أرقام استبعاد المحدي في الجدولي? السابقي?، ثم التجربة الميدانية مع مستخدمي? صم التي ما تزال خطوتنا التالية المعلنة.

9-4الصعوبات وحدود النظام

واجهتنا أربع صعوبات رئيسة تستحق التوثيق لأ? أي فريق يعيد بناء نظام مشابه سيصطدم بها. الأولى والأخطر: اكتشفنا بالتحقق م? نقاط معصم الجسم أ? مقاطع KArSL مسجلة بصورة مرآتية (بسبب طبيعة تسجيل Kinect)، بينما كاميرا الهاتف ترسل إطارا غير مرآتي؛ أي أ? النموذج كا? سيستقبل م? المستخدمي? الحقيقيي? إشارات معكوسة الاتجاه ع? كل ما تدرب عليه. الحل كا? برمجيا بسيطا بعد فهم السبب: يعكس المتصفح الإطار قبل التتبع، فل حاجة لإعادة التدريب. والثانية انعكاس يمي?/يسار مماثل في أوضاع اليد الواحدة على كاميرا الحاسوب، إذ يسمي المتعقِّب اليد الظاهرة «يسرى» في إطار مرآتي وهي يمنى المستخدم، فصحح التعيي?. والثالثة تخزي? عامل الخدمة الذي كسر وعد العمل دو ? اتصال في أول نشر كما روينا في الفصل الثالث. والرابعة حصص تحميل البيانات التي فرضت علينا بناء مجموعة التدريب م? مصدري? غير متكافئي الجودة.

أما حدود النظام الحالية فنلخصها@مانة: المفردات 261 إشارة لا تغطي محادثة مفتوحة؛ والدقة على أشخاص جدد (76.5%) تعني خ@ً في نحو ربع الإشارات، يخفف منه عرض التخمينات البديلة وزر الحذف السريع لكنه لا يلغيه؛ ونقاط الوجه في خصائصنا لم تُدرَّب فعليا على نحو الوجه كالنفي والاستفهام فتبقى تعبيرات المستخدم النحوية خارج فهم النموذج حاليا؛ والحل الشامل ثقيل على الهواتف الضعيفة لم نقِس أداءه بعد على أجهزة منخفضة المواصفات؛ وفوق ذلك كله لم يُختبَر النظام بعد مع مستخدمي? صم حقيقيي?، وهي التجربة التي ستقرر وحدها أسئلة مثل راحة التقطيع بخفض اليدي? وملءمة عتبة الثقة. هذه الحدود ليست اعتراف هزيمة؛ إنها خريطة الطريق التي رسمناها للفصل الأخير م? هذا الكتاب.

الخاتمة والفاق المستقبلية

الخاتمة

انطلقنا م? مشهد واحد: يد ترتفع في قاعة انتظار فل يفهمها أحد. وانتهينا بنظام يعمل على هاتف عادي، يرى اليد ويفهم إشارتها وينطق جملتها، دو? أ? يرسل صورة واحدة إلى أي خادم.

ما بنيناه تحديدا: خط معالجة كامل داخل المتصفح م? الكاميرا إلى الصوت؛ وثلثة أوضاع تعرف تتدرج م? قواعد هندسية بل تدريب إلى نموذج عصبي بمفردات م? 261 إشارة؛ ومنهجية تقييم صارمة باستبعاد مجدٍّ كامل أعطتنا 85.8% على مئة إشارة و76.5% على المفردات الأوسع؛ وقاموس مرئي وإمكانيـــة تعليم محليـــة تجعل? النظام ينمو بيد مستخدميه. وأهم ما تعلمناه لم يك? رقما: كا? أ? أصـــعب أخطاء هذا النوع م? الأنظمة لا تسك? النموذج بل ما حوله، م? كاميرا مرآتية في البيانات، إلى يمي? ويسار متبادلي?، إلى عامل خدمة يكسر وعدا قطعناه للمستخدم.

الفاق والتطويرات المستقبلية

أول خطواتنا التالية معلنة ومحددة: تجربة ميدانية مع مستخدمي? صم حقيقيي?،@سئلة الراحة والدقة اليومية لا يجيب عنها مختبر. وبالتوازي، جمع بيانات محلية برضا أصحابها وإشرافهم، تُحفظ نقاطا رقمية لا صورا، لتعليم النموذج لهجات الإشـــارة المحلية التي لا تغطيها قواعد البيانات الحالية.

وعلى صعيد النموذج، يعيد خط أدواتنا إنتاج نموذج أوسع متى تـــوفرت مقـــاطع KArSL الخام كاملة، والطريق مفتوح بعده نحو مفردات المجموعة الكاملة (502 إشارة) بالأدوات ذاتها. وتبقى جبهتا? نعرفهما جيدا: إدخال نحو الوجـــه، كـــالنفي والاستفهام، في الفهم فعليا لا في الخصائص فقط، وقياس الأداء على هواتف منخفضة المواصفات وتخفيف الحل الشامل لها. فإذا تحقق نصف هذه الخطة صار نظامنا أداة

محادثة يومية لا أداة مواقف فقط؛ وهذا هو الطموح كما ينبغي أ? يكو ?: محددا وقابلً للقياس.

الملحق

الشارات السهلة التسع وقواعد أدائها

يعرض هذا الملحق الإشارات التسع في وضع الإشارات السهلة كما هي معرفة في قاموس التطبيق وقواعد قراءتها البرمجية، مع العبارة التي ينطقها النظام لكل منها. يستطيع المستخدم إعادة تعيين؟ أي م؟ هذه العبارات م؟ داخل التطبيق.

الشارة	العبارة المنطوقة	كيفية الداء
مرحبا	مرحبا	كف مفتوحة ثابتة، الأصابع للعلى
مع السلمة	مع السلمة	التلويح بكف مفتوحة يمينا ويسارا، وتُكشف م؟ الحركة لا م؟ شكل ثابت
أحبك	أحبك	الإبهام والسبابة والخنصر مفتوحة
نعم	نعم	الإبهام للعلى
لا	لا	الإبهام للسفل
تمام	تمام	الإبهام يلمس السبابة، والباقي مفتوح
شكرا	شكرا	السبابة والوسطى مفتوحتا؟ ومتباعدتا؟
لحظة م؟ فضلك	لحظة م؟ فضلك	السبابة وحدها للعلى
اتصل بي	اتصل بي	الإبهام والخنصر مفتوحا؟

تُقرأ هذه الأشكال بقواعد على حالة الأصابع الخمس كما فصّلنا في الفصل الرابع، ولا تحتاج إلى أي تدريب، وتبقى ثابتة ما لم يغيّرها المستخدم م؟ إعدادات التطبيق أو يعلّمه إشارات إضافية خاصة به.

المراجع

[1] World Health Organization, World Report on Hearing. Geneva, Switzerland: World Health Organization, 2021.

[2] United Nations, "International Day of Sign Languages." [Online]. Available: https://www.un.org/en/observances/sign-languages-day

[3] United Nations, Convention on the Rights of Persons with Disabilities. New York, NY, USA: United Nations, 2006.

[4] A. A. Sidig, H. Luqman, S. Mahmoud, and M. Mohandes, "KArSL: Arabic Sign Language Database," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 20, no. 1, pp. 1–19, 2021, doi: 10.1145/3423420.

[5] G. Latif, N. Mohammad, J. Alghazo, R. AlKhalaf, and R. AlKhalaf, "ArASL: Arabic Alphabets Sign Language Dataset," Data in Brief, vol. 23, art. 103777, 2019, doi: 10.1016/j.dib.2019.103777.

[6] H. Luqman, "An Efficient Two-Stream Network for Isolated Sign Language Recognition Using Accumulative Video Motion," IEEE Access, vol. 10, pp. 93785–93798, 2022.

[7] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-Device Real-Time Hand Tracking," arXiv:2006.10214, 2020.

[8] C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," arXiv:1906.08172, 2019.

[9] I. Grishchenko and V. Bazarevsky, "MediaPipe Holistic — Simultaneous Face, Hand and Pose Prediction, on Device," Google Research Blog, Dec. 2020. [Online]. Available: https://research.google/blog/mediapipe-holistic-simultaneous-face-hand-and-pose-prediction-on-device/

[10] Vite, "Vite Documentation." [Online]. Available: https://vite.dev

[11] Mozilla, "MediaDevices.getUserMedia()," MDN Web Docs. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/MediaDevices/getUserMedia

[12] Google, "Hand Landmarker Guide," MediaPipe Solutions Documentation. [Online]. Available: https://developers.google.com/mediapipe/solutions/vision/hand_landmarker

[13] ONNX, "Open Neural Network Exchange." [Online]. Available: https://onnx.ai

[14] Microsoft, "ONNX Runtime Web," ONNX Runtime Documentation. [Online]. Available: https://onnxruntime.ai/docs/get-started/with-javascript.html

[15] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019.

[16] Rhasspy, "Piper: A Fast, Local Neural Text to Speech System," GitHub. [Online]. Available: https://github.com/rhasspy/piper

[17] ElevenLabs, "Text to Speech," ElevenLabs Documentation. [Online]. Available: https://elevenlabs.io/docs

[18] Mozilla, "Web Speech API," MDN Web Docs. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API

[19] Cloudflare, "Cloudflare Pages Documentation." [Online]. Available: https://developers.cloudflare.com/pages/

[20] Mozilla, "Service Worker API," MDN Web Docs. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/Service_Worker_API