كلية الهندسة
قسم الذكاء الصنعي وعلوم البيانات

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

أعدت هذه الأطروحة

لإنجاز مقرر المشروع الفصلي في اختصاص الذكاء الصنعي وعلوم البيانات

# Voice Command Recognition

إعداد الطلاب:

رنيم ربيع  أيهم السالم

أشراف:

د.ميساء أبو قاسم  م.وسام السحلي

**الملخص:**

**يتناول هذا المشروع تصميم وتنفيذ نظام ذكي للتحكم في جهاز الكومبيوتر باستخدام الأوامر الصوتية، بهدف**

**توفير وسيلة تفاعلية أكثر سهولة ومرونة في الاستخدام، خاصة لذوي الاحتياجات الخاصة أو في بيئات**

**تتطلب التشغيل بدون استخدام اليدين. يعتمد النظام على التعرف التلقائي على الصوت لتحويل الأوامر**

**. المنطوقة إلى إجراءات مباشرة يتم تنفيذها على نظام التشغيل**

**يتم استقبال الصوت من خلال ميكروفون، ثم معالجته باستخدام أدوات برمجية مثل مكتبة**

**، لترجمة الأوامر إلى تعليمات تنفذ مهام مثل فتح Python بلغة PyAuto و TensorFlow البرامج، التحكم**

**. في الوسائط، أو تنفيذ أوامر النظام المختلفة**

**يمثل المشروع خطوة نحو دمج تقنيات الذكاء الاصطناعي بالتفاعل الإنساني مع الحاسوب، ويوفر أساسًا**

**لتطوير أنظمة أكثر تطورًا في المستقبل تدعم اللغات الطبيعية بشكل أوسع وتتيح تخصيص الأوامر حسب**

**احتياجات المستخد م**

**Abstract:**

**This project focuses on the design and implementation of an intelligent system for controlling a computer using voice commands. The goal is to provide a more flexible and user-friendly interaction method, especially beneficial for individuals with physical disabilities or in hands-free operation environments. The system relies on speech recognition to convert spoken commands into executable actions on the operating system.Audio input is captured through a microphone and processed using programming tools such as the SpeechRecognition and PyAutoGUI libraries in Python. These tools interpret the commands and translate them into instructions that perform tasks such as opening applications, controlling media playback, browsing, or executing various system functions.**

**The project represents a step towards integrating artificial intelligence into human-computer interaction and lays the foundation for more advanced systems that support natural language processing and customizable command sets in the future**

## المحتويات

قائمة بأهم المصطلحات :

| المصطلح التقني | الاختصار | الوصف |
|---|---|---|
| Whisper | | نموذج لتحويل الكلام الصوتي إلى نص واستخراج ميزات صوتية عميقة |
| ASR (Automatic Speech Recognition) | | تحويل الصوت إلى نص تلقائيًا |
| Whisper Transcription | | النص الناتج عن تفريغ الصوت بواسطة Whisper |
| Whisper Audio Features | | تمثيل رقمي عالي المستوى للصوت مستخرج من Whisper |
| Intent Recognition | | تحديد نية المستخدم من الأمر الصوتي |
| Action–Object Pair | | تمثيل النية بصيغة فعل ← كائن |
| Feature Extraction | | تحويل البيانات الخام إلى ميزات رقمية قابلة للتعلم |
| Feature Fusion | | دمج ميزات صوتية ونصية ضمن تمثيل واحد |
| TF-IDF | | تمثيل نصي إحصائي يعتمد على أهمية الكلمات |
| SBERT | | نموذج يحول الجمل إلى متجهات دلالية تمثل المعنى |
| Semantic Embeddings | | تمثيل رقمي يعكس المعنى بدل الكلمات الحرفية |
| Logistic Regression (LR) | | نموذج تصنيف إحصائي بسيط وفعّال |
| SVD (TruncatedSVD) | | تقليل أبعاد الميزات مع الحفاظ على المعلومات المهمة |
| Dimensionality Reduction | | تقليل عدد الأبعاد لتحسين الكفاءة والتعميم |
| MFCC | | ميزات صوتية تقليدية تمثل الطيف الترددي |
| Delta / Delta-Delta | | ميزات تمثل التغير الزمني للصوت |
| PCA | | تقنية لتقليل الأبعاد أو تصور البيانات |
| t-SNE | | طريقة لتصور التداخل أو الفصل بين الفئات |

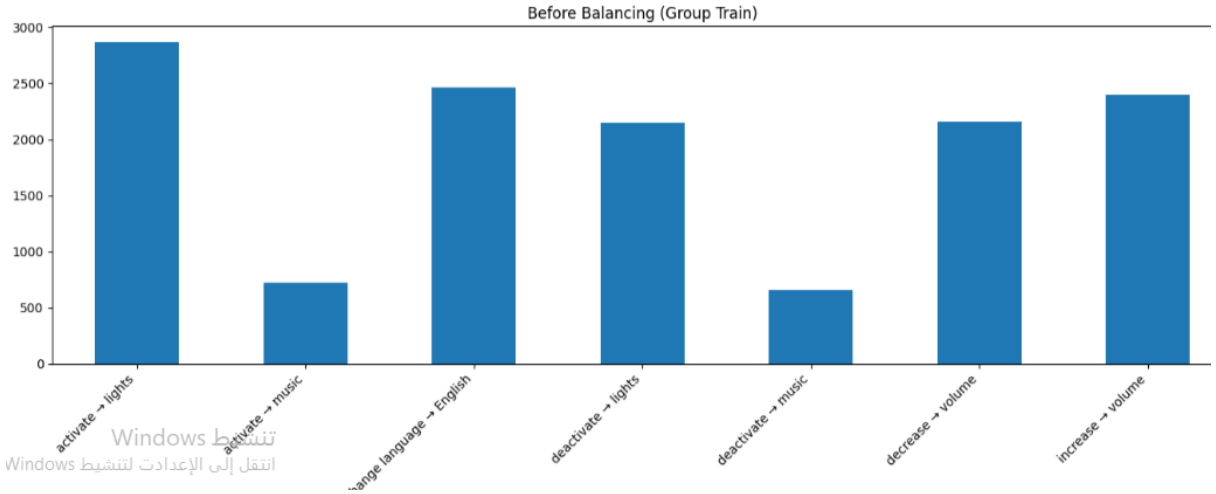
| | | |
|---|---|---|
| Data Augmentation | | توليد عينات صوتية إضافية لتحسين التعميم |
| Time Stretching | | تغيير سرعة الصوت دون تغيير نبرته |
| Pitch Shift | | تغيير طبقة الصوت |
| Noise Injection | | إضافة ضجيج لمحاكاة بيئات واقعية |
| Speaker ID | | معرف المتحدث المستخدم لفصل البيانات |
| Group Split | | تقسيم يمنع تسرب المتحدثين بين المجموعات |
| Stratified Split | | تقسيم يحافظ على توزيع الفئات |
| Data Leakage | | تسرب معلومات الاختبار إلى التدريب |
| Overfitting | | حفظ بيانات التدريب بدل تعلم أنماط عامة |
| Regularization | | تقنيات لتقليل فرط التخصيص |
| Dropout | | تعطيل عشوائي للوحدات العصبية لمنع الحفظ الزائد |
| Confidence Threshold | | عتبة تحدد قبول أو رفض التنبؤ |
| Margin (Confidence Gap) | | الفرق بين أعلى احتمالين لتقدير الثقة |
| Rejection Mechanism | | آلية لرفض الأوامر غير الواضحة أو خارج النطاق |
| Out-of-Domain (OOD) | | أوامر خارج نطاق الأوامر المدعومة |
| In-Domain Ambiguous | | أوامر ضمن المجال لكنها غير مباشرة أو غير واضحة |
| Confusion Matrix | | جدول يوضح الأخطاء بين الفئات |
| Classification Report | | تقرير يعرض Precision وRecall وF1 لكل فئة |
| Accuracy | | نسبة التنبؤات الصحيحة إجمالًا |
| Precision | | نسبة التنبؤات الصحيحة داخل فئة معينة |
| Recall | | نسبة الحالات الصحيحة التي تم اكتشافها |

| | | |
|---|---|---|
| F1 Score | | مقياس يوازن بين Precisionو Recall |
| Macro F1 | | متوسط F1 عبر جميع الفئات بالتساوي |
| Quantitative Evaluation | | تقييم رقمي باستخدام مقاييس إحصائية |
| Qualitative Evaluation | | تقييم عملي عبر أمثلة واقعية |
| Real-World Evaluation Dataset | | تسجيلات واقعية لاختبار الأداء العملي |
| ASR Errors | | أخطاء ناتجة عن تفريغ Whisper |
| Linguistic Noise | | ضجيج لغوي ناتج عن أخطاء النص |
| Semantic Confusion | | التباس بين أوامر متقاربة دلاليًا |
| Model Generalization | | قدرة النموذج على العمل على بيانات غير مرئية |
| Learning Curves | | منحنيات توضح تحسن الأداء مع زيادة البيانات |
| Practical Performance Gap | | الفرق بين النتائج الرقمية والأداء الواقعي |
| Confidence-Based Decision | | اتخاذ القرار بناءً على مستوى الثقة |
| Automatic Speech Recognition | ASR | نظام تحويل الكلام المنطوق إلى نص مكتوب، ويُعد المرحلة الأساسية في معظم أنظمة فهم الأوامر الصوتية |
| Spoken Language Understanding | SLU | مجال يركّز على فهم نية المستخدم ومعنى الأمر المنطوق، ويتضمن مهام مثل تصنيف النية واستخراج الكيانات |
| Intent Classification | IC | مهمة تحديد النية الأساسية من الأمر الصوتي (مثل تشغيل، إيقاف، زيادة) |
| Slot Filling | SF | مهمة استخراج الكيانات أو التفاصيل المرتبطة بالأمر مثل الكائن أو الموقع (music ، lights ،volume) |
| End-to-End SLU | E2E-SLU | نظام يقوم بتنفيذ فهم اللغة المنطوقة مباشرة من الصوت دون فصل صريح بين ASRو NLU |
| Large Language Model | LLM | نموذج لغوي ضخم مُدرَّب على كميات هائلة من النصوص ويُستخدم لفهم اللغة أو توليدها |

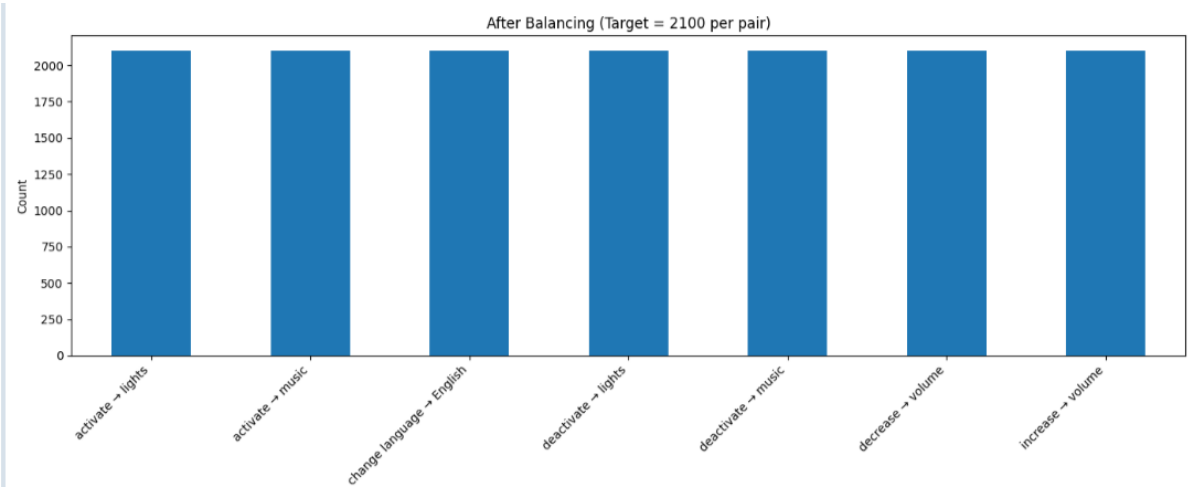
| | | |
|---|---|---|
| Audio-Language Model | — | نموذج يجمع بين تمثيلات صوتية ولغوية ضمن إطار واحد لفهم الكلام |
| SpeechVerse | — | متعددة SLU لتنفيذ مهام LLM نموذج صوت–لغة واسع النطاق يدمج الصوت مع |
| Task Fine-Tuning | Task-FT | ضبط النموذج المسبق التدريب على مهمة محددة مثل تصنيف النية أو استخراج الخانات |
| Word Confusion Network | WCN | يأخذ بعين الاعتبار عدة بدائل للكلمات بدل نص واحد ASR تمثيل احتمالي لمخرجات فقط |
| Robust SLU | — | مصممة لتحمّل أخطاء التفريغ الصوتي والضوضاء SLU أنظمة |
| Zero-shot Learning | ZS | قدرة النموذج على تنفيذ مهمة دون تدريب مسبق على بيانات من نفس المجال |
| Prompting | — | بدل إعادة تدريبه (Prompts) أسلوب توجيه النموذج عبر تعليمات نصية |
| QA-driven SLU | — | إلى صيغة سؤال–جواب لتمكين الفهم بدون بيانات موسومة SLU تحويل مهمة |
| ZS-Whisper-SLU | — | End-to-End بأسلوب SLU لتنفيذ Whisper يعتمد على Zero-shot نظام |
| Perfect Parsing | PP | مقياس يعبّر عن نسبة الأوامر التي تم فهم نيتها وخاناتها بالكامل دون خطأ |
| Cross-Corpus Evaluation | — | تقييم النموذج على بيانات من مجال مختلف عن بيانات التدريب لقياس التعميم |
| Data Augmentation | — | توسيع بيانات التدريب عبر توليد عينات إضافية (حقيقية أو اصطناعية) |
| LLM-generated Data | — | بيانات تدريب يتم توليدها باستخدام نماذج لغوية كبيرة لتحسين أداء المصنّفات |
| WHISMA | — | Zero-shot بصيغة SLU واحد لتنفيذ LLM مع Whisper إطار يدمج |
| SLU-GLUE | — | مجموعة مهام معيارية لتقييم أنظمة فهم اللغة المنطوقة |
| SCoT | — | متقدّم لتحسين الاستدلال خطوة بخطوة Prompting أسلوب |
| MR (Multi-Reasoning) | MR | LLM تقنية لتعزيز التفكير متعدد الخطوات داخل |
| Multi-modal Learning | — | تعلم يعتمد على دمج أكثر من وسيط (صوت، نص، فيديو) |
| Multi-modal Intent Recognition | MIR | مهمة تحديد النية بالاعتماد على عدة وسائط بدل الصوت فقط |
| LGSRR | — | نموذج متعدد الوسائط لتمييز النية يعتمد على علاقات دلالية بين الوسائط |
| Weighted F1-score | WF1 | مقياس أداء يأخذ بعين الاعتبار عدم توازن الفئات |

| | | |
|---|---|---|
| Word Error Rate | WER | مقياس لقياس أخطاء التفريغ الصوتي في ASR |
| Exact Match | EM | نسبة العينات التي تم التنبؤ بها بشكل مطابق تمامًا للتصنيف الصحيح |
| Confidence Threshold | — | عتبة احتمالية تُستخدم لقبول أو رفض التنبؤ |
| Rejection Mechanism | — | آلية تمنع تنفيذ أوامر خارج نطاق النظام أو ذات ثقة منخفضة |

# قائمة بالاشكال:




Figure( 6)




figure(7)

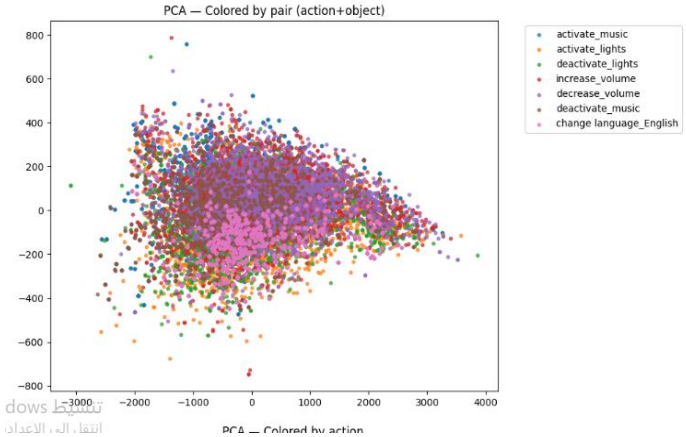



Figure 1

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| activate → lights | 0.9860 | 0.9643 | 0.9750 | 364 |
| activate → music | 0.8947 | 0.9659 | 0.9290 | 88 |
| change language → English | 0.9928 | 0.9928 | 0.9928 | 278 |
| deactivate → lights | 0.9810 | 0.9663 | 0.9736 | 267 |
| deactivate → music | 0.9383 | 0.9744 | 0.9560 | 78 |
| decrease → volume | 0.9699 | 0.9663 | 0.9681 | 267 |
| increase → volume | 0.9590 | 0.9690 | 0.9640 | 290 |
| accuracy | | | 0.9712 | 1632 |
| macro avg | 0.9602 | 0.9713 | 0.9655 | 1632 |
| weighted avg | 0.9717 | 0.9712 | 0.9713 | 1632 |

Figure 2

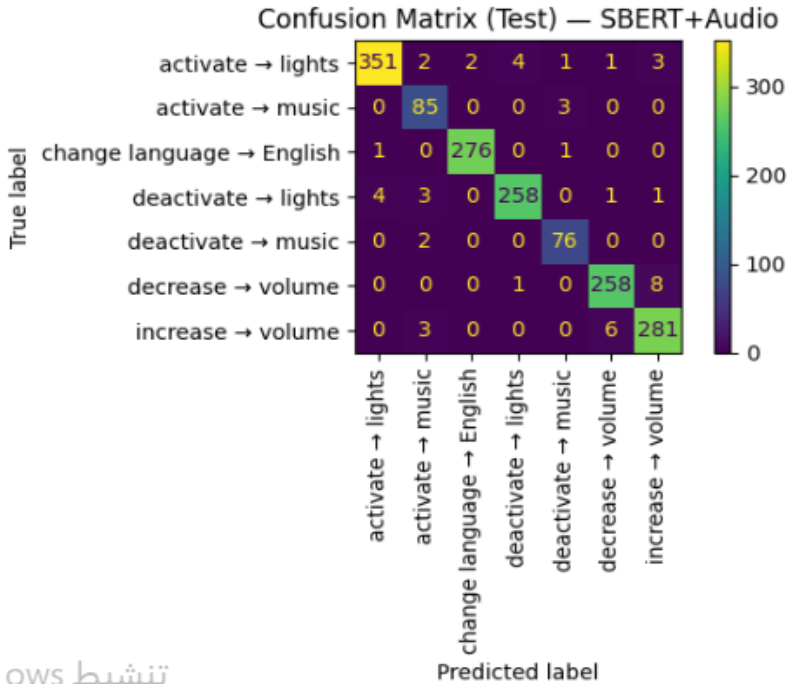



Figure 3

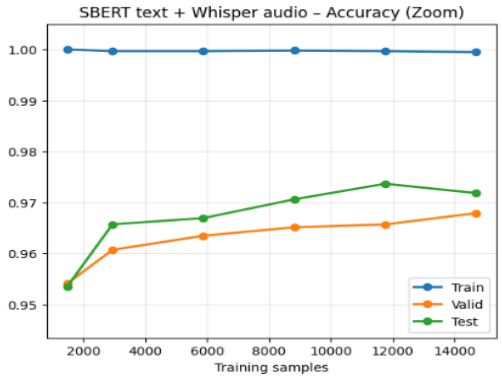



Figure5

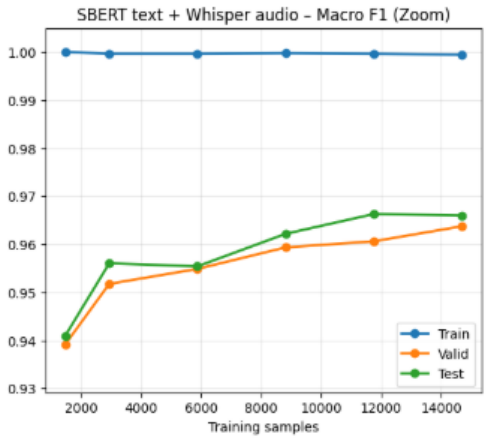



Figure4

## الفصل الأول:

### 1.1مقدمة عن المشروع :

شهد التفاعل بين الإنسان والحاسوب تطورًا ملحوظًا في السنوات الأخيرة مع التقدم السريع في مجالات الذكاء الاصطناعي ومعالجة الإشارات الصوتية. ومن بين أهم هذه التطورات أنظمة التحكم الصوتي التي تتيح للمستخدم تنفيذ الأوامر والتحكم بالأجهزة باستخدام الصوت بدلاً من وسائل الإدخال التقليدية مثل لوحة المفاتيح أو الفأرة. يركّز المشروع على فهم الأمر الصوتي وظيفيًا وليس فقط تحويل الصوت إلى نص، مما يجعله أكثر ملاءمة للتطبيقات العملية مثل أنظمة التحكم بالحاسوب، المنازل الذكية، والمساعدات الصوتية.

### 1.2 الهدف من المشروع :

يهدف هذا المشروع إلى تحقيق مجموعة من الأهداف، من أبرزها:

1. تصميم نظام ذكي للتعرف على الأوامر الصوتية اعتمادًا على الصوت الطبيعي للمستخدم.
2. استخراج النص من الصوت باستخدام نموذج Whisper ASR ومعالجة أخطاء التحويل الصوتي.
3. تصنيف الأوامر الصوتية إلى فئات واضحة *(Action → Object)* بدل الاكتفاء بالنص الخام.

4. بناء نموذج Fusion يجمع بين:
    - الميزات الصوتية المستخرجة منWhisper
    - الميزات النصية المستخرجة من نصWhisper
5. معالجة مشكلة عدم توازن البيانات باستخدام تقنيات Data Augmentation و Downsampling.
6. تحقيق دقة عالية في تصنيف الأوامر مع الحفاظ على قابلية التعميم.
7. تمهيد الطريق لإضافة مرحلة Reasoningلاحقًا باستخدام LLM أو قواعد منطقية

## 1.3 المشكلة التي يقوم المشروع بحلها

تواجه طرق التفاعل التقليدية مع أجهزة الحواسيب، مثل استخدام لوحة المفاتيح والفأرة، العديد من التحديات التي تؤثر على كفاءة المستخدمين وراحتهم ومن بين أبرز هذه التحديات، الصعوبات التي يواجهها ذوو الاحتياجات الخاصة، خصوصًا أولئك الذين يعانون من إعاقات حركية تمنعهم من استخدام الوسائل التقليدية للتحكم بالجهاز. في هذا السياق، تسعى أنظمة التحكم الصوتي إلى تقديم بديل عملي يمكنهم من التعامل مع الكومبيوتر بسهولة واستقلالية أكبر

إلى جانب ذلك، تسهم هذه الأنظمة في تحسين الإنتاجية، حيث تتيح للمستخدمين تنفيذ الأوامر بشكل أسرع، دون الحاجة إلى التنقل بين القوائم أو الضغط على أزرار متعددة كما تلعب دورًا مهمًا في تقليل الجهد البدني والإجهاد الناتج عن الاستخدام الطويل للأدوات التقليدية، مما يساعد على تجنب بعض المشاكل الصحية مثل آلام المعصم أو الكتف. وفي بيئات العمل متعددة المهام أو في الظروف التي تكون فيها اليدان مشغولتين، يصبح التحكم الصوتي وسيلة مثالية للتفاعل مع الحاسوب، إذ يوفّر للمستخدم مرونة أكبر ويجعل تجربة الاستخدام أكثر سلاسة وفعالية.

## الفصل الثاني :

## 2.1الدراسة النظرية:

### Whisper

يُعد نموذج Whisperمن النماذج المتقدمة في مجال التعرف التلقائي على الكلام (Automatic Speech Recognition – ASR)، وهو نموذج عميق قائم على بنية Transformer Encoder–Decoderتم تدريبه على نطاق واسع باستخدام بيانات صوتية متعددة اللغات ومتعددة المجالات، مما يمنحه قدرة عالية على التعميم والتعامل مع تنوع اللهجات، الضجيج، وسياقات الكلام المختلفة. تعتمد معمارية Whisper على مبدأ تحويل الإشارة الصوتية الخام إلى تمثيل طيفي مضغوط عبر حساب Mel-Spectrogram، حيث يتم تقسيم الإشارة الزمنية إلى نوافذ قصيرة، ثم تحويلها إلى المجال الترددي باستخدام تحويل فورييه السريع، وبعدها إسقاطها على مقياس Mel الذي يعكس الحساسية السمعية البشرية للترددات. يتم تمرير هذا التمثيل الطيفي إلى Encoderالذي يتعلم تمثيلات صوتية عالية المستوى عبر طبقات الانتباه الذاتي(Self-Attention) ، مما يسمح للنموذج بفهم العلاقات الزمنية الطويلة بين المقاطع الصوتية والتقاط الأنماط الصوتية الدلالية مثل النبرة، الإيقاع، والمحتوى الفونيمي. بعد ذلك، يقوم Decoderبتوليد النص المقابل خطوة بخطوة من خلال آلية الانتباه المتقاطع (Cross-Attention) التي تربط التمثيلات الصوتية المخرجة من الـ Encoder بالتسلسل النصي الناتج، بحيث يتم التنبؤ بالكلمة التالية اعتماداً على السياق الصوتي والسياق اللغوي السابق في آنٍ واحد. من الناحية العلمية، لا يقتصر Whisper على كونه نظام تفريغ صوتي تقليدي، بل يعمل كنظام نمذجة احتمالية لتطابق الإشارة الصوتية مع اللغة المكتوبة، حيث يتعلم التوزيع الاحتمالي المشترك بين الصوت والنص، مما يتيح له تصحيح أخطاء النطق، التعامل مع التوقفات، واستنتاج الكلمات غير الواضحة اعتماداً على السياق العام للجملة. إضافة إلى إنتاج النص، يستطيع Whisper استخراج تمثيلات عددية عميقة Audio) (Embeddingsتمثل جوهر الإشارة الصوتية بشكل مضغوط، وهي متجهات رقمية تحتوي معلومات مركّبة عن الخصائص الطيفية، الإيقاع الزمني، البنية الفونيمية، والنمط الصوتي العام، ويمكن إعادة استخدامها في مهام أخرى مثل تصنيف الأوامر الصوتية، التعرف على المتحدث، أو فهم النوايا. تتميز هذه الميزات بأنها عالية المستوى وغير يدوية، أي أنها لا تعتمد على مؤشرات تقليدية مثل MFCC فقط، بل تُستخرج بشكل تعلّمي مباشر من البيانات، مما يجعلها أكثر قدرة على تمثيل المعنى الصوتي الفعلي بدلاً من الاكتفاء بالخصائص الفيزيائية الخام. وبذلك، يشكّل Whisper نظاماً متكاملاً يجمع بين التمثيل الطيفي، الفهم الزمني العميق، والنمذجة اللغوية الاحتمالية، الأمر الذي يجعله مناسباً ليس فقط لتحويل الصوت إلى نص، بل أيضاً كمصدر غني لاستخراج ميزات صوتية

دلالية متقدمة تُستخدم في التطبيقات الذكية مثل أنظمة الأوامر الصوتية، المساعدات الذكية، وتحليل الكلام المتقدم.

### Sentence-BERT (SBERT)

يُعد نموذج Sentence-BERT (SBERT) امتداداً متقدماً لبنية BERT التقليدية، وقد تم تصميمه خصيصاً لمعالجة أحد القيود الأساسية في نماذج المحولات اللغوية، والمتمثل في عدم كفاءة تمثيل الجمل كاملةً بشكل متجهات دلالية قابلة للمقارنة المباشرة. يعتمد SBERT على بنية Siamese Transformer Network حيث يتم تمرير جمل أو نصوص متعددة عبر مُشفّر (Encoder) مشترك المعاملات لاستخراج تمثيلات دلالية متناسقة ضمن نفس الفضاء المتجهي، مما يسمح بقياس التشابه الدلالي بين النصوص باستخدام مقاييس بسيطة مثل Cosine Similarity بدلاً من الحاجة إلى تمرير الأزواج النصية كاملةً عبر النموذج في كل مرة. من الناحية العلمية، يتعلم SBERT تمثيلات نصية عميقة عبر التدريب على مهام مثل الاستدلال اللغوي الطبيعي (Natural Language Inference) ومطابقة الأزواج الدلالية، مما يجعله قادراً على التقاط المعنى السياقي الكامل للجملة بدلاً من تمثيل الكلمات منفردة. عند تطبيقه في مشروع معالجة الأوامر الصوتية، يتم استخدام SBERT لتحويل النص الناتج عن Whisper إلى متجهات دلالية كثيفة (Dense Semantic Embeddings) تمثل النية والمعنى المجرد للأمر بدلاً من الاعتماد على تطابق الكلمات السطحية فقط، وهو ما يتيح للنظام التعامل مع إعادة الصياغة، المرادفات، والاختلافات اللغوية بشكل أكثر مرونة. تمر عملية استخراج الميزات النصية عبر عدة مراحل تبدأ بتنظيف النص وتوحيده، ثم تمريره إلى SBERT الذي يُنتج تمثيلاً عدديّاً ثابت الأبعاد يعكس العلاقات النحوية، المعنى السياقي، والبنية الدلالية العامة للجملة، حيث تُستخدم هذه المتجهات لاحقاً في مهام التصنيف أو الدمج متعدد الوسائط مع الميزات الصوتية. تتميز الميزات المستخرجة من SBERT بأنها عالية المستوى ودلالية بطبيعتها، إذ لا تمثل فقط تواتر الكلمات كما في الأساليب الإحصائية التقليدية مثل TF-IDF ، بل تعبّر عن المقصود الحقيقي للنص حتى في حال اختلاف الصياغة اللفظية، مما يعزز قدرة النظام على تعميم الفهم وتحسين دقة تصنيف النوايا. وبذلك، يشكل SBERT طبقة دلالية متقدمة في خط أنابيب المشروع، حيث يعمل كجسر بين النص الخام والمعنى التجريدي، ويساهم في رفع موثوقية النظام عند تفسير الأوامر النصية المشتقة من الصوت.

## Logistic Regression

يُعد Logistic Regression من الخوارزميات الإحصائية الشائعة في تعلم الآلة، ويُستخدم بشكل رئيسي في مهام التصنيف لتقدير احتمال انتماء العينة إلى فئة معينة اعتمادًا على مجموعة من الميزات العددية. يقوم النموذج بحساب تركيب خطي بين الميزات المدخلة ومعاملات قابلة للتعلّم، ثم تمرير الناتج عبر دالة لوجستية لتحويله إلى قيمة احتمالية محصورة بين 0 و1. في حالة التصنيف متعدد الفئات، يتم استخدام دالة Softmax لإنتاج توزيع احتمالي عبر جميع الفئات الممكنة، مما يسمح للنموذج باتخاذ قرار نهائي بناءً على أعلى احتمال متوقَّع.

يتم تدريب النموذج عبر تقليل دالة الخسارة اللوجستية باستخدام خوارزميات تحسين عددية مثل LBFGS أو SAGA لضمان تقارب مستقر وسريع. كما يمكن تطبيق تقنيات التنظيم مثل L2 Regularization للحد من فرط التكيّف عبر تقليل تأثير الأوزان الكبيرة، مما يساعد على تحسين قدرة النموذج على التعميم على بيانات جديدة. يتم التحكم بدرجة التنظيم من خلال المعاملC ، حيث تمثل القيم الأصغر تنظيمًا أقوى والقيم الأكبر مرونة أعلى للنموذج.

يمتاز Logistic Regression ببساطته الحسابية وسرعة تدريبه وسهولة تفسيره مقارنةً بالنماذج العميقة، مما يجعله مناسبًا عند العمل مع تمثيلات مضغوطة مثل الميزات الصوتية والنصية منخفضة الأبعاد. ولهذا السبب، تم اعتماده في هذا المشروع كنموذج تصنيف نهائي لتحديد نية الأوامر اعتمادًا على الميزات متعددة الوسائط المدمجة، نظرًا لتوازنه بين الأداء والاستقرار وقابلية التفسير.

## TruncatedSVD (Dimensionality Reduction)

تُستخدم تقنية TruncatedSVD كأداة لتقليل الأبعاد (Dimensionality Reduction) في البيانات عالية الأبعاد من خلال تحليل البنية الخطية للمصفوفات واستخراج المكونات الأكثر تمثيلاً للمعلومات الأساسية. تعتمد هذه الطريقة على تفكيك المصفوفة الأصلية إلى عدد محدود من المركبات الخطية (Singular Vectors) التي تفسّر أكبر قدر ممكن من التباين في البيانات، مما يسمح بتمثيل كل عينة بعدد أقل من الأبعاد مع الحفاظ على أهم الخصائص الإحصائية. في هذا المشروع، تم تطبيق TruncatedSVD على الميزات الصوتية والنصية بهدف تقليل التعقيد الحسابي، وتحسين كفاءة التعلم، والحد من الضوضاء والارتباطات الزائدة بين الأبعاد، مما يساعد النموذج اللاحق على التعلّم من تمثيل مضغوط وأكثر استقرارًا.

### TF-IDF (Text Representation)

تم استخدام تقنية TF-IDF (Term Frequency–Inverse Document Frequency) لتمثيل النصوص بشكل عددي يعكس أهمية الكلمات داخل المستند مقارنةً ببقية مجموعة البيانات. تعتمد هذه الطريقة على مبدأ إعطاء وزن أعلى للكلمات التي تظهر بشكل متكرر داخل نص معين ولكنها أقل شيوعًا عبر بقية النصوص، مما يساعد على إبراز الكلمات الأكثر دلالة دلاليًا للتمييز بين الأوامر المختلفة. بعد تحويل النصوص إلى متجهات TF-IDF عالية الأبعاد، تم تطبيق تقنيات تقليل الأبعاد مثل TruncatedSVD لاستخراج تمثيلات نصية مضغوطة تحتفظ بالمعلومات الأساسية دون الاحتفاظ بالضجيج اللغوي أو التكرار غير المفيد. يتيح هذا الأسلوب بناء تمثيل عددي فعّال للنصوص يمكن دمجه لاحقًا مع الميزات الصوتية ضمن إطار تعلم متعدد الوسائط.

### Rejection Mechanism (Decision Layer)

تم تضمين آلية رفض (Rejection Mechanism) ضمن طبقة القرار (Decision Layer) بهدف تحسين موثوقية النظام وتقليل التنبؤات غير الدقيقة في الحالات غير الواضحة أو خارج نطاق المجال (Out-of-Domain). تعتمد هذه الآلية على تحليل مخرجات النموذج الاحتمالية (Prediction Probabilities) وتطبيق شروط قائمة على العتبات (Thresholds) مثل الحد الأدنى لاحتمال الفئة المتوقعة، والفارق بين أعلى احتمالين(Confidence Margin) ، وذلك لتحديد ما إذا كان التنبؤ موثوقًا بدرجة كافية أم يجب رفضه. بالإضافة إلى ذلك، يتم دمج قواعد قائمة على معالجة النص (Text Normalization) واكتشاف النوايا غير المؤكدة (Ambiguous In-Domain Intents) للتعامل مع أخطاء التفريغ الصوتي (ASR Errors) أو الصياغات غير القياسية. تساهم هذه الآلية في جعل النظام أكثر واقعية وأمانًا من خلال تقليل القرارات الخاطئة وإجبار النموذج على الامتناع عن التنبؤ عندما تكون درجة الثقة منخفضة.

## الفصل الثالث: الدراسة المرجعية

### 3.1 مقدمة

شهد مجال فهم اللغة المنطوقة (Spoken Language Understanding – SLU) تطورًا كبيرًا في السنوات الأخيرة نتيجة التقدم في تقنيات الذكاء الاصطناعي، ومعالجة الإشارات الصوتية، والنماذج اللغوية العميقة. وقد ركزت الأبحاث الحديثة على تحسين قدرة الأنظمة الصوتية على فهم نية المستخدم بدقة، والتعامل مع أخطاء التعرف التلقائي على الكلام، بالإضافة إلى دمج أكثر من وسيط (الصوت، النص، السياق) لتحقيق أداء أفضل في البيئات الواقعية.

### 3.2 ا الدراسة الأولى

تتناول دراسة SpeechVerse: A Large-scale Generalizable Audio-Language Model (2024) تقديم نموذج موحّد لمعالجة اللغة المنطوقة يُعرف باسم SpeechVerse، يدمج بين التمثيل الصوتي العميق والنمذجة اللغوية ضمن بنية واحدة، بهدف فهم الكلام المنطوق بشكل مباشر دون الفصل التقليدي بين مرحلتي التعرف على الكلام (ASR) وفهم النص. اعتمد النموذج على بيانات صوتية واسعة ومتنوعة تغطي لهجات وبيئات مختلفة، وتم تدريبه ضمن إطار متعدد المهام يشمل التعرف على الكلام، تصنيف النية، واستخراج الخانات .(Slot Filling) أظهرت النتائج أن نموذج SpeechVerseبعد الضبط المخصص للمهمة (Task-Finetuned) حقق دقة تقارب 84–85%في مهمة تصنيف النية على مجموعة بيانات SLURP، إلى جانب أداء تنافسي في مهام أخرى مرتبطة بفهم اللغة المنطوقة. ومع ذلك، بيّنت الدراسة أن مستوى الأداء يختلف باختلاف نوع المهمة وإعدادات الضبط، مما يعكس طبيعة النماذج متعددة المهام وصعوبة الاعتماد على مقياس واحد يعبّر بدقة عن الأداء الكلي للنموذج

### 3.3الدراسة الثانية

تناقش دراسة ASR-Robust Spoken Language Understanding via Word Confusion Networks (2024) أثر أخطاء التفريغ الصوتي على أداء أنظمة فهم اللغة المنطوقة المعتمدة على نماذج لغوية كبيرة. أظهرت النتائج أن دقة تصنيف النية باستخدام نماذج مثل GPT-3.5 تنخفض من حوالي 88.1%عند استخدام النص الصحيح إلى نحو 84.6%عند الاعتماد على النص الناتج عنASR ، مما يوضح حساسية هذه النماذج لأخطاء التفريغ. كما بيّنت الدراسة أن استخدام شبكات ارتباك الكلمات (WCN) لتحسين المتانة أمام أخطاء ASR لم ينجح في استعادة الأداء الكامل، مما يؤكد أن جودة التفريغ الصوتي تبقى عاملًا حاسمًا في أداء أنظمة SLU الواقعيةASR .

### 3.5 الدراسة الثالثة

تناولت دراسة “Prompting Whisper for QA-driven Zero-shot End-to-End Spoken Language Understanding” (2024)استخدام نموذج Whisper ضمن إطار QA-driven Zero-shot End-to-End Spoken Language Understandingبهدف تنفيذ مهام فهم اللغة المنطوقة دون الاعتماد على الفصل التقليدي بين التعرف على الصوت (ASR) وتحليل النص. تم تقييم النموذج عبر اختبارات Cross-Corpusعلى مجموعات بيانات مختلفة مثل FSC وSmartLight، حيث أظهر نموذج ZS-Whisper-SLUقدرة عالية على التعميم. حقق النموذج دقة تصنيف نية بلغت نحو 95%على مجموعةFSC ، كما وصل إلى Accuracy ≈ 91.6% و SLU-F1 ≈ 90.9% على مجموعةSmartLight ، مع معدلات Perfect Parsingتجاوزت .82%وتشير هذه

النتائج إلى أنWhisper ، عند توجيهه بأسلوب مناسب، قادر على تحقيق أداء تنافسي في مهام SLU حتى في غياب بيانات تدريب مخصصة من نفس المجال، مما يبرز إمكاناته كنظام End-to-End لفهم الأوامر الصوتية.

### 3.6 الدراسة الرابعة

تناولت دراسة WHISMA: A Speech-LLM to Perform Zero-Shot Spoken Language Understanding (2024) تطوير نموذج يدمج بين Whisperونموذج لغوي كبير (LLM) لتنفيذ مهام فهم اللغة المنطوقة بأسلوب Zero-Shotدون تدريب مخصص على بيانات SLU. تم تقييم النموذج على مهام متعددة ضمن إطار SLU-GLUE، حيث حقق نموذج WHISMA متوسط دقة يقارب 72%، وارتفع الأداء إلى نحو 79%عند استخدام تقنيات تعزيز الاستدلال مثل Self-CoTوMulti-Reasoning، متفوقًا على بعض النماذج الصوتية–اللغوية الأخرى. تشير النتائج إلى أن دمج Whisper مع LLM قادر على تعميم فهم النية الصوتية عبر مهام مختلفة دون بيانات تدريب مباشرة، رغم أن الأداء لا يزال أدنى من الأنظمة الإشرافية المتخصصة، مما يبرز إمكاناته كنموذج مرن وقابل للتعميم في تطبيقات SLU الحديثة.

### 3.8 الدراسة الخامسة

Multi-modal Intent Recognition (MIR) – EMNLP 2025

تناولت هذه الدراسة تطوير نموذج متعدد الوسائط لفهم نية المستخدم من خلال دمج المعلومات الصوتية والبصرية والنصية ضمن إطار موحد، أُطلق عليه اسم LGSRR.تم تقييم النموذج على مجموعتي بيانات قياسيتين هما MIntRec2.0وIEMOCAP-DA، حيث أظهر تفوقًا واضحًا مقارنةً بالنماذج متعددة الوسائط السابقة ونماذج اللغة الكبيرة العامة. حقق النموذج دقة تصنيف بلغت حوالي 60.5%على مجموعةMIntRec2.0 ، وقرابة 74.9%على مجموعةIEMOCAP-DA ، مع تحسن ملحوظ في مقاييس F1. كما أظهرت دراسات الإزالة أن كل مكوّن في النموذج يساهم بشكل فعّال في تحسين الأداء، مما يؤكد أهمية التصميم المتخصص بدل الاعتماد المباشر على نماذج LLM العامة. تعكس هذه النتائج فاعلية الدمج المنهجي بين الوسائط المختلفة في تحسين فهم النية في البيئات الواقعية المعقدة.

| الرقم | عنوان الدراسة | سنة النشر | الفكرة الأساسية | المنهجية / النموذج المستخدم | مجموعات البيانات | معايير التقييم | أبرز النتائج القابلة للاعتماد |
|---|---|---|---|---|---|---|---|
| 1 | **SpeechVerse: A Large-scale Generalizable Audio-Language Model** | 2024 | تطوير نموذج صوت–لغة موحّد لفهم اللغة المنطوقة مباشرة دون فصل تقليدي بين ASR وNLU | Audio Encoder + LLM، تدريب متعدد المهام (ASR، IC، SF)مع Fine-tuning مخصص | SLURP، LibriSpeech، VoxPopuli | Intent Accuracy، SLU-F1، WER | حقق نموذج SpeechVerse Task-Finetunedدقة تقارب **84–85%** في تصنيف النية علىSLURP ، مع أداء تنافسي في مهام Slot FillingوASR، وأظهر أن الأداء يختلف باختلاف نوع المهمة وآلية الضبط |
| 2 | **ASR-Robust SLU via Word Confusion Networks** | 2024 | تحسين متانة أنظمة SLU ضد أخطاء التفريغ الصوتي عبر تمثيل احتمالي للنص | ASR → Word Confusion Network (WCN) → LLM | ATIS، NMSQA | Accuracy، F1، Exact Match | أظهر استخدام WCN تحسنًا ملحوظًا في F1 و EMمقارنة باستخدام ناتج ASR أحادي (1-best)، خصوصًا في سيناريوهات الضوضاء والأخطاء الصوتية |
| 3 | **Prompting Whisper for QA-driven Zero-shot End-to-End SLU** | 2024 | استخدام Whisperكنظام End-to-End لفهم الأوامر الصوتية بأسلوب Zero-shot | Whisper + QA-style Prompting (ZS-Whisper-SLU) | FSC، SmartLight، SLURP | Accuracy، SLU-F1، Perfect Parsing | حقق النموذج **≈95% Accuracy**علىFSC ، و **Accuracy ≈ 91.6%**و **SLU-F1 ≈ 90.9%**على SmartLight، مع PP تجاوزت 82%، مما يثبت قدرة Whisper على التعميم دون تدريب مخصص |
| 4 | **WHISMA: A Speech-LLM for Zero-shot Spoken Language Understanding** | 2024 | دمج Whisper مع LLM واحد ضمن إطار موحّد لتنفيذ SLU بدون بيانات تدريب من نفس المجال | Whisper + LLaMA-3 + Prompting (WHISMA) | SLU-GLUE (SST-2, QQP, QNLI, RTE, SciTail) | Accuracy (Zero-shot) | حقق النموذج متوسط دقة **≈79%**عبر مهام SLU-GLUE، مع تفوق واضح عند استخدام تقنيات Prompting محسّنة (SCoT + MR) مقارنة بالأنظمة المعيارية |
| 5 | **Multi-modal Intent Recognition (MIR)** | 2025 | تحسين التعرف على النية عبر دمج الصوت والنص والفيديو ضمن نموذج متعدد الوسائط | Audio + Text + Visual Encodersمع نموذجLGSRR | MIntRec2.0، IEMOCAP-DA | Accuracy، F1، Recall، Weighted F1 | تفوق LGSRR على النماذج الأحادية الوسيط، محققًا **ACC ≈ 60.5% (MIntRec2.0)** و **ACC ≈ 74.9% (IEMOCAP-DA)**، مما يثبت فاعلية الدمج متعدد الوسائط |

# الفصل الرابع: منهجية العمل

## 4.1 مقدمة

تُعد أنظمة التعرف على الأوامر الصوتية من أكثر تطبيقات الذكاء الاصطناعي تحدّيًا، وذلك بسبب الطبيعة المعقدة للإشارات الصوتية التي تتأثر بعوامل متعددة مثل الضوضاء، اختلاف اللهجات، سرعة النطق، ونبرة الصوت.

يتطلب فهم الأوامر الصوتية بدقة الجمع بين تحليل الإشارة الصوتية ومعالجة اللغة الطبيعية، حيث إن الاعتماد على الصوت فقط أو النص فقط قد يؤدي إلى ضعف في التعميم عند العمل في بيئات واقعية.

يساهم تطوير أنظمة فعّالة للتعرف على الأوامر الصوتية في تحسين التفاعل بين الإنسان والحاسوب، خاصة في تطبيقات التحكم بالأجهزة، الأنظمة الذكية، ودعم المستخدمين ذوي الاحتياجات الخاصة.

في هذا الفصل، يتم استعراض المنهجية المتبعة لبناء نظام ذكي قادر على التعرف على الأوامر الصوتية وتنفيذها، بدءًا من جمع البيانات الصوتية، مرورًا بمرحلة التحضير والمعالجة، وصولًا إلى بناء نموذج الدمج (Fusion) وتقييم أدائه باستخدام مقاييس علمية دقيقة.

## 4.2 منهجية العمل العامة

تم بناء النظام المقترح باتباع منهجية متكاملة شملت:

جمع البيانات الصوتية والنصية ،إعداد البيانات وتحسين جودتها، توحيد الأوامر وربطها بالنية الصحيحة(Mapping) ، توسيع البيانات (Augmentation) ثم موازنتها، استخراج الميزات الصوتية والنصية ، بناء نموذج دمج بين الصوت والنص، تقييم النموذج باستخدام بيانات اختبار واقعية. اعتمدت المنهجية على التحليل العملي والتجريبي مع مقارنة النتائج في كل مرحلة، لضمان بناء نموذج موثوق وقابل للتعميم

## 4.3 مجموعة البيانات المستخدمة

تُعد مجموعة بيانات Fluent Speech Commands (FSC)من مجموعات البيانات المرجعية في مجال فهم الأوامر الصوتية(Spoken Language Understanding) ، حيث تحتوي على تسجيلات صوتية لـ 97متحدثًا قاموا بنطق 248عبارة مختلفة تمثل 31نية (Intent) فريدة، ويتم تمثيل كل أمر ضمن ثلاث خانات رئيسية (Slots) هي : الفعل(Action) ، الكائن(Object) ، والموقع(Location) ، بهدف توفير معيار (Benchmark)لتقييم نماذج الفهم الدلالي للأوامر الصوتية من البداية إلى النهاية. تم جمع البيانات باستخدام الحشد الجماعي (Crowdsourcing) من مشاركين في الولايات المتحدة وكندا، حيث طُلب من كل مشارك نطق كل عبارة مرتين بترتيب عشوائي، مع الحصول على موافقة صريحة لمشاركة التسجيلات الصوتية إلى جانب معلومات ديموغرافية مجهولة الهوية مثل القدرة اللغوية، اللغة الأولى، اللغة المستخدمة في العمل أو الدراسة، الجنس، والفئة العمرية. خضعت التسجيلات الصوتية لاحقًا لعملية تحقق جودة مستقلة من قبل مجموعة أخرى من المقيّمين، وتم استبعاد أي ملفات تحتوي على ضجيج مرتفع، أو صوت غير واضح، أو أخطاء في العبارة المنطوقة. يتم تنظيم البيانات ضمن مجلدات بحسب معرّف المتحدث(Speaker ID) ، بحيث يحتوي كل مجلد على ملفات صوتية بصيغة WAVبمعدل أخذ عينات 16كيلوهرتز وقناة واحدة، إضافة إلى ملفات توصيف بصيغة CSV لكل من مجموعات التدريب والتحقق والاختبار، تتضمن معلومات تفصيلية لكل عينة مثل مسار الملف الصوتي، معرّف المتحدث، النص المنطوق، والفئات الدلالية المرتبطة به، حيث تشمل قيم Action أوامر مثل (change language, activate, deactivate, increase, decrease, bring)، وقيم Objectعناصر مثل (music, lights, volume, lamp, newspaper, juice, socks, shoes, English, German)وغيرها، مما يجعل هذه المجموعة موردًا غنيًا ومناسبًا لتدريب وتقييم أنظمة الأوامر الصوتية القائمة على النية

## 4.4 تحضير البيانات(Data Preparation)

### 4.4.1 حذف الأعمدة غير الضرورية من مجموعات البيانات

تم في هذه المرحلة تنقية مجموعات البيانات الخاصة بالتدريب والتحقق والاختبار من الأعمدة غير الضرورية التي لا تدخل ضمن نطاق مهمة فهم الأوامر الصوتية. شمل ذلك استبعاد معلومات الموقع، لكونها غير مستخدمة في تمثيل النية ضمن هذا المشروع، إضافةً إلى إزالة الأعمدة الوصفية الزائدة الناتجة عن عمليات حفظ أو تصدير البيانات. تهدف هذه الخطوة إلى تبسيط بنية البيانات والاحتفاظ فقط بالسمات المرتبطة مباشرة بالمحتوى الدلالي

والصوتي، مما يساهم في تقليل التعقيد وتحسين وضوح البيانات قبل مراحل المعالجة اللاحقة.

### 4.4.2 إنشاء المسارات الكاملة للملفات الصوتية

تم في هذه المرحلة توحيد مسارات الملفات الصوتية من خلال ربط المسار النسبي لكل ملف مع المسار الأساسي لمجلد البيانات، وذلك بهدف إنشاء مسار كامل وموحّد لكل عينة صوتية. نتج عن هذه العملية إضافة عمود جديد ضمن إطار البيانات يحتوي على المسارات الكاملة للملفات الصوتية، مما يسهّل الوصول إليها واستخدامها لاحقًا في مراحل المعالجة المسبقة واستخراج الميزات. كما تم التحقق من صحة هذه المسارات عبر استعراض عينة منها للتأكد من جاهزيتها للاستخدام.

### 4.4.3 تحويل النصوص المفرّغة إلى أوامر معيارية

تم اعتماد آلية لمعالجة النصوص المفرّغة من الصوت (Transcription Processing) بهدف ربط كل نص بنية الأمر المقابلة له. تعتمد هذه الآلية على توحيد صيغة النص عبر تحويله إلى أحرف صغيرة lowercase، ثم تحليل محتواه للكشف عن كلمات مفتاحية مرتبطة بمجالات محددة مثل التحكم بالإضاءة، تغيير اللغة، مستوى الصوت، أو تشغيل الموسيقى. بناءً على هذه الكلمات، يتم تحديد الفعل (Action) والكائن (Object) اللذين يعكسان المعنى المقصود من الأمر. وفي الحالات التي لا يتطابق فيها النص مع القواعد المحددة، يتم الإبقاء على القيم الأصلية دون تعديل.

### 4.4.4 تحديث حقول الأوامر استناداً إلى نتائج المعالجة النصية

تم في هذه المرحلة اعتماد آلية ربط دلالي (Command Mapping) تهدف إلى تحويل الأوامر الفردية أو الكلمات المفتاحية المباشرة إلى تمثيل موحّد بصيغة (Action–Object Pair). تعتمد هذه الآلية على إسناد معنى تنفيذي واضح لكل أمر بسيط مثل *stop* أو *pause* أو *play*، بحيث يتم ربطه بالفعل المناسب) مثل deactivate أو activate) والكائن المرتبط به) مثل music أو volume) ، بغضّ النظر عن كونه جزءًا من جملة كاملة أو أمرًا منفردًا.

يساهم هذا الأسلوب في توحيد تمثيل الأوامر داخل النظام، وتقليل الغموض الدلالي، وتمكين نموذج التصنيف من التعامل مع الأوامر القصيرة والمباشرة بنفس الكفاءة التي يتعامل بها مع الأوامر الأطول، دون الحاجة إلى تحليل لغوي معقّد.

### 4.4.5 تصفية الصفوف المرتبطة بكائنات غير مستهدفة

يتم تعريف قائمة unwanted_objects التي تحتوي على مجموعة من القيم النصية المرتبطة بكائنات غير مرغوبة. بعد ذلك، يتم تصفية إطار البيانات df لإزالة الصفوف

التي يكون فيها العمود object منتمياً إلى هذه القائمة، ثم يتم إعادة ضبط الفهرس باستخدام reset_index لضمان الحفاظ على تسلسل الصفوف بعد عملية الحذف

### 4.4.6 إنشاء تسمية مركبة للأوامر وعرض توزيعها الإحصائي

يتم إنشاء عمود جديد باسم pair عبر دمج قيمتي action وobject في سلسلة نصية واحدة بحيث يمثل هذا العمود الصيغة الموحدة للأمر.

### 4.4.7 تقسيم البيانات باستخدام Group Split بناءً على المتحدث ( Speaker-Based Splitting)

يتم استخدام GroupShuffleSplit من مكتبة Scikit-learn لتنفيذ عملية تقسيم البيانات مع مراعاة معرف المتحدث speakerId كعامل تجميع رئيسي. يتم أولاً التحقق من وجود العمود speakerId داخل إطار البيانات لضمان توفر معلومات المتحدث قبل تنفيذ التقسيم. يتم تقسيم البيانات إلى مجموعة تدريب بنسبة 80% ومجموعة مؤقتة بنسبة 20% بحيث لا يظهر أي متحدث في كلا المجموعتين في الوقت نفسه. لاحقاً، يتم تقسيم المجموعة المؤقتة إلى مجموعتي تحقق (Validation) واختبار (Test) بنسبة متساوية باستخدام نفس آلية الفصل حسب المتحدث.

### 4.4.8 تحميل بيانات التدريب المخصّصة ومعالجة الLabels

يتم تحديد مجلد يحتوي على ملفات التقسيم المحفوظة مسبقاً، ثم تحميل ملف التدريب باستخدام مكتبة Pandas. بعد تحميل البيانات، يتم إنشاء عمود pair عبر دمج قيم action وobject في تمثيل نصي موحد، كما يتم إضافة عمود جديد باسم is_aug لتحديد ما إذا كانت العينة أصلية أو ناتجة عن عمليات تعزيز لاحقة، ويتم ضبط قيمته الابتدائية إلى صفر لجميع العينات الأصلية. بعد ذلك، يتم حساب عدد التكرارات لكل فئة ضمن العمود pair وعرض توزيعها قبل تنفيذ أي تعديل على البيانات.

### 4.4.9 تعريف دالة تعزيز الإشارة الصوتية (Audio Augmentation)

تم تطبيق آلية لتعزيز الإشارة الصوتية بهدف زيادة تنوّع بيانات التدريب وتحسين قدرة النموذج على التعميم. تعتمد هذه الآلية على توحيد سعة الإشارة لمنع التشبع الرقمي، ثم تطبيق مجموعة من التحويلات العشوائية الخفيفة على الصوت، تشمل الإزاحة الزمنية للإشارة(Temporal Shift) ، وتغيير سرعة النطق(Time Stretching) ، وتعديل طبقة الصوت(Pitch Shift) ، بالإضافة إلى إضافة ضجيج منخفض الشدة (Noise Injection).
تُطبَّق هذه التحويلات باحتمالات متفاوتة بحيث تحافظ على المعنى العام للأمر الصوتي، مع إدخال تنوّع واقعي يحاكي اختلاف ظروف التسجيل.

### 4.4.10 توليد عينات إضافية للوصول إلى عدد مستهدف لكل فئة

من أجل معالجة عدم التوازن بين فئات الأوامر، تم اعتماد استراتيجية توليد عينات إضافية للفئات الأقل تمثيلاً وصولًا إلى عدد مستهدف موحّد لكل فئة. عند انخفاض عدد العينات ضمن فئة معيّنة عن العدد المطلوب، يتم توليد عينات جديدة من نفس الفئة باستخدام تقنيات تعزيز الإشارة الصوتية، مع الحفاظ على الانتماء الدلالي للفئة. يهدف هذا الإجراء إلى تقليل تحيّز النموذج نحو الفئات الأكثر تكرارًا دون التأثير على التوزيع الدلالي العام للبيانات.

### 4.4.11 دمج العينات الأصلية مع العينات المعززة

بعد الانتهاء من عملية التعزيز، تم دمج العينات الأصلية مع العينات المعزَّزة ضمن مجموعة تدريب واحدة موحّدة. يضمن هذا الدمج الحفاظ على جميع العينات ضمن إطار بيانات متكامل، مما يسمح بالتعامل معها بشكل متسق خلال مراحل التدريب اللاحقة، مثل إعادة التوازن أو التقسيم أو التقييم.

### 4.4.12 تقليل العينات الزائدة للوصول إلى حجم موحّد لكل فئة

يتم تجميع البيانات المدمجة وفق العمود pair، ثم فصل العينات الأصلية عن العينات المعززة استناداً إلى قيمة العمود is_aug. في حال كان العدد الإجمالي لعينات فئة معينة أقل من أو يساوي العدد المستهدف، يتم الاحتفاظ بجميع العينات دون حذف. أما في حال تجاوز العدد الإجمالي القيمة المستهدفة، يتم اختيار عينات للاحتفاظ بها بحيث تُعطى الأولوية للعينات الأصلية، ويتم استكمال العدد المطلوب بعينات معززة عند الحاجة. بعد تحديد العينات النهائية لكل فئة، يتم دمج جميع الأجزاء في إطار بيانات موحّد وإعادة ترتيب الصفوف عشوائياً لضمان توزيع متوازن.

### 4.4.13 عرض التوزيع الإحصائي للفئات قبل وبعد الموازنة

يتم استخدام مكتبة Matplotlib لرسم مخططين عموديين يوضّحان توزيع عدد العينات لكل فئة ضمن العمود pair قبل تنفيذ عملية الموازنة وبعدها كما في الشكل (6) (7).

## 4.5 استخراج الميزات الصوتية (Feature Extraction)

### 4.5.1 تحميل نموذج Whisper والمعالج المرافق له

يتم تحميل نموذج Whisper من Hugging Face باستخدام الإصدار whisper-small عبر WhisperForConditionalGeneration، إلى جانب تحميل المعالج WhisperProcessor المسؤول عن تحويل الإشارة الصوتية إلى تمثيل طيفي مناسب للإدخال إلى النموذج. يتم تحديد جهاز التنفيذ تلقائياً اعتماداً على توفر وحدة معالجة

رسومية CUDA أو المعالجة عبر CPU، ثم نقل النموذج إلى الجهاز المحدد وضبطه على وضع التقييم (evaluation mode) لمنع تحديث الأوزان أثناء الاستخراج. أظهرت إعدادات النموذج أن البعد الداخلي لمتجهات التمثيل الصوتي ( Embedding Dimension أو d_model) يساوي 768، وهو ما يمثل طول المتجه الناتج لكل ملف صوتي بعد استخراج الميزات من المشفر (Encoder).

### 4.5.2 تحميل الملفات الصوتية وتوحيدها إلى صيغة Mono بمعدل kHz16

تُعرَّف دالة لتحميل الملفات الصوتية باستخدام مكتبة Librosa، حيث يتم تحويل كل ملف إلى قناة واحدة وتوحيد معدل أخذ العينات عند 16,000ثم تحويل الإشارة الناتجة إلى Tensor بصيغة float32 لاستخدامها في إطار PyTorch. بعد تنفيذ هذه العملية على مجموعات البيانات، بلغت أبعاد التمثيل الصوتي الخام (14700, 773) لمجموعة التدريب، و(1806, 773) لمجموعة التحقق، و(1632, 773) لمجموعة الاختبار، بما يعكس عدد العينات وطول التمثيل الزمني لكل ملف صوتي بعد المعالجة الأولية.

### 4.5.3 تنفيذ التجميع المتوسط مع مراعاة القناع الزمني (Masked Mean Pooling)

تم اعتماد أسلوب Masked Mean Poolingلاستخلاص تمثيل صوتي ثابت الطول من المخرجات الزمنية لمشفّر Whisper. تعتمد هذه الآلية على حساب المتوسط الزمني للمتجهات الخفية مع مراعاة Attention Maskعند توفره، بحيث يتم تجاهل الإطارات الناتجة عن الحشو (Padding) والتركيز فقط على الإطارات الفعلية التي تمثل المحتوى الصوتي الحقيقي. يساهم هذا الأسلوب في تحسين دقة التمثيل النهائي وتقليل تأثير الحشو، مع ضمان الاستقرار العددي أثناء الحساب. وفي الحالات التي لا يتوفر فيها القناع الزمني أو لا يكون متوافقًا مع أبعاد المخرجات، يتم اللجوء إلى المتوسط الزمني التقليدي كخيار بديل لضمان استمرارية عملية الاستخراج.

### 4.5.4 استخراج ميزات Whisper باستخدام إعدادات إدخال محددة

تم استخراج الميزات الصوتية باستخدام نموذج Whisper عبر تمرير الإشارات الصوتية بإعدادات إدخال موحّدة تضمن اتساق التمثيل بين جميع العينات. شملت هذه الإعدادات توحيد معدل العيّنة (Sampling Rate = 16 kHz) وتطبيق الحشو الزمني (Padding) لضبط أطوال الإشارات المختلفة. ينتج عن ذلك تمثيل طيفي على شكل Mel-based Input Featuresيتم تمريره إلى مشفّر Whisper لاستخلاص متجهات خفية عالية المستوى (Hidden Representations) تعبّر عن الخصائص الصوتية الدلالية. تم

اعتماد طول إدخال أدنى موحّد لضمان ثبات أبعاد الميزات بين العينات، مع تعطيل عمليات التدريب أثناء الاستخراج لضمان كفاءة الحساب والتركيز على التمثيل فقط.

### 4.5.5 تجميع المتجهات الصوتية وتحويلها إلى تمثيل ثابت الطول

بعد استخراج المتجهات الخفية من المشفر، يتم تطبيق التجميع الزمني عبر حساب المتوسط على البعد الزمني باستخدام masked mean pooling عند توفر قناع الانتباه، أو المتوسط التقليدي عند عدم توفره. ينتج عن هذه العملية متجه تمثيلي ثابت الطول لكل ملف صوتي بطول 768 بُعداً، ثم يتم تجميع جميع المتجهات الناتجة في مصفوفة واحدة باستخدام np.vstack، مع إنشاء أسماء أعمدة رقمية متسلسلة من الشكل f0 إلى f767 لتمثيل الأبعاد العددية للميزات.

## 4.6 استخراج النص باستخدام Whisper (Whisper ASR Transcription)

### 4.6.1 تجهيز مجلدات التخزين المؤقت والحفظ النهائي

يتم في هذا الجزء تحديد مجلدين رئيسيين هما مجلد التخزين المؤقت CACHE_DIR لتجميع نتائج التفريغ النصي، ومجلد الحفظ النهائي SAVE_DIR لتخزين ملفات الMeta data بعد إضافة النصوص (تمثل ملفات الـ Metadata في هذا المشروع جداول وصفية تحتوي على النصوص المستخرجة من نموذج Whisper إلى جانب الLabels المرتبطة بكل عينة، والمسارات الصوتية، والمعلومات التعريفية الأخرى. وتُستخدم هذه الملفات كحلقة وصل بين البيانات الصوتية الخام والميزات العددية المستخرجة لاحقاً، مما يتيح مواءمة النصوص مع التمثيلات الصوتية أثناء مراحل التدريب والتقييم)

### 4.6.2 تحميل ملفات الMeta data والتحقق من أبعاد التقسيمات

تم تحميل ملفات الMeta data الخاصة بمجموعات التدريب والتحقق والاختبار من المسارات المحددة، حيث بلغ حجم مجموعة التدريب 14700 عينة مع تسعة أعمدة بيانات، بينما بلغ حجم مجموعة التحقق 1806 عينات بثمانية أعمدة، ومجموعة الاختبار 1632 عينة بثمانية أعمدة. بعد التحميل، تم التحقق من وجود العمود pair داخل كل مجموعة، وإنشاؤه عند الحاجة عبر دمج قيم action وobject في تمثيل نصي موحد لضمان اتساق الLabels بين جميع التقسيمات.

### 4.6.3 تحميل نموذج Whisper وتنفيذ التفريغ الصوتي

تم تحميل نموذج Whisper باستخدام الإصدار small لتنفيذ عملية تحويل الصوت إلى نص. بعد ذلك، تم تنفيذ عملية التفريغ النصي على مجموعات البيانات الثلاث بشكل

متتابع، حيث استغرقت عملية تفريغ مجموعة التدريب زمناً يقارب 53 دقيقة و50 ثانية لمعالجة 14700 ملف صوتي، بمعدل يقارب 4.55 ملفات في الثانية. كما استغرقت عملية تفريغ مجموعة التحقق حوالي 8 دقائق و20 ثانية لمعالجة 1806 ملفات، بينما استغرقت مجموعة الاختبار حوالي 7 دقائق و24 ثانية لمعالجة 1632 ملفاً صوتياً.

### 4.6.4 إضافة النصوص المفرّغة وإدارة cache أثناء التنفيذ

تمت إضافة عمود جديد باسم whisper_text إلى كل مجموعة بيانات، حيث تم توليد النصوص المفرّغة لكل ملف صوتي مع الاعتماد على آلية كاش مبنية على تجزئة مسار الملف لتخزين النتائج النصية ومنع إعادة الحساب عند التشغيل المستقبلي. خلال التنفيذ الحالي، لم تكن هناك نتائج مخزنة مسبقاً في cache، حيث بلغت قيمة العناصر المسترجعة من cache 0 في جميع المجموعات. في مجموعة التدريب، تم تسجيل 2815 ملفاً صوتياً مفقوداً لم يكن لها مسار صالح، بينما لم تُسجل أي ملفات مفقودة في مجموعتي التحقق والاختبار، كما لم يتم تسجيل أي حالات فشل في عملية التفريغ عبر جميع المجموعات.

### 4.6.5 أبعاد البيانات بعد إضافة whisper_text

بعد الانتهاء من عملية التفريغ، أصبحت مجموعة التدريب تحتوي على 14700 صفاً و10 أعمدة بعد إضافة عمود النص المفرّغ، في حين أصبحت مجموعة التحقق تحتوي على 1806 صفوف و9 أعمدة، ومجموعة الاختبار تحتوي على 1632 صفاً و9 أعمدة. تم التأكد من نجاح إضافة العمود الجديد إلى جميع التقسيمات قبل الانتقال إلى مراحل تمثيل النصوص واستخراج الميزات الدلالية.

## 4.7 تحضير البيانات ومواءمة الميزات ( Data Preparation and Feature Alignment)

### 4.7.1 تحميل الMeta data والميزات الصوتية والتحقق من الأبعاد

تم تحميل ملفات الMeta data الخاصة بمجموعات التدريب والتحقق والاختبار، حيث بلغت الأبعاد (14700, 10) لمجموعة التدريب، و(1806, 9) لمجموعة التحقق، و(1632, 9) لمجموعة الاختبار بعد تضمين النصوص الناتجة عن Whisper والLabels التعريفية. كما تم تحميل ملفات الميزات الصوتية المستخرجة مسبقاً (Audio Features)، وبلغت الأبعاد الخام لهذه الميزات (14700, 773) لمجموعة التدريب، و(1806, 773) لمجموعة التحقق، و(1632, 773) لمجموعة

الاختبار، بما يعكس عدد العينات وعدد الأعمدة العددية المرتبطة بالتمثيل الصوتي (Audio Embeddings).

### 4.7.2 توحيد النصوص والLabels وتجهيز الحقول الأساسية

تم التأكد من توفر العمود whisper_text ضمن ملفات الMeta data، مع معالجة القيم المفقودة وتحويلها إلى صيغة نصية موحّدة، كما تم توحيد أعمدة action وobject والتأكد من وجود العمود pair أو إنشائه عند الحاجة لتمثيل وسم الأمر بصيغة نصية موحدة (Label Normalization). بعد ذلك، تم استخراج قوائم النصوص لكل تقسيم، مع عرض مثال لنص مفرّغ مثل "stop music." للتحقق من سلامة محتوى البيانات النصية (Text Data).

### 4.7.3 تنظيف الميزات الصوتية وإعادة ضبط الفهارس

تم حذف الأعمدة غير الضرورية من ملفات الميزات الصوتية مثل العمود is_aug في حال وجوده، ثم إعادة ضبط الفهارس في كل من ملفات الMeta data وملفات الميزات الصوتية لضمان تطابق ترتيب العينات بين المصدرين ( Index Alignment). بعد ذلك، تم التحقق من تطابق عدد الصفوف بين الملفات المرتبطة بكل تقسيم لمنع أي عدم توافق أثناء عملية الدمج (Misalignment Prevention).

### 4.7.4 دمج الMeta data مع الميزات الصوتية وبناء مجموعات البيانات النهائية

تم دمج الأعمدة النصية والLabels من ملفات الMeta data مع الأعمدة الرقمية من ملفات الميزات الصوتية باستخدام الدمج الأفقي (Horizontal Concatenation)، مما نتج عنه مجموعات بيانات نهائية بأبعاد (14700, 774) لمجموعة التدريب، و(1806, 774) لمجموعة التحقق، و(1632, 774) لمجموعة الاختبار، حيث تحتوي هذه المجموعات على النصوص، الLabels، والميزات العددية ضمن بنية موحدة (Unified Dataset Structure).

### 4.7.5 إعداد الLabels النهائية وقوائم النصوص لكل تقسيم

تم استخراج الLabels النهائية من العمود pair وتحويلها إلى مصفوفات نصية بأبعاد (14700,) لمجموعة التدريب، و(1806,) لمجموعة التحقق، و(1632,) لمجموعة

النصوص الناتجة عن Whisper من العمود whisper_text وتحويلها إلى قوائم نصية منفصلة لكل تقسيم (Text Lists)، مع الاحتفاظ بعينات نصية مرجعية لمتابعة جودة البيانات.

### 4.7.6 توحيد وترتيب الأعمدة الرقمية للميزات الصوتية

تم استخراج الأعمدة الرقمية من مجموعات البيانات الثلاث، حيث بلغ عدد الأعمدة الرقمية 768 عموداً في كل تقسيم (Feature Dimensions)، ثم تم تحديد مجموعة الأعمدة المشتركة بينها والتأكد من تطابقها بالكامل بعدد 768 بُعداً عددياً. بعد ذلك، تم ترتيب هذه الأعمدة ترتيباً رقمياً موحداً وفق أسماء الميزات مثل f0 و f1 و f2 لضمان ثبات ترتيب الأبعاد عبر جميع المجموعات (Feature Ordering).

### 4.7.7 إنشاء مصفوفات الميزات الصوتية بعد المواءمة

بعد توحيد الأعمدة الرقمية وترتيبها، تم تحويل الميزات الصوتية لكل تقسيم إلى مصفوفات NumPy تمثل الإدخال العددي النهائي للنماذج اللاحقة ( Numeric Feature Matrices)، حيث بلغت أبعاد هذه المصفوفات (14700, 768) لمجموعة التدريب، و(1806, 768) لمجموعة التحقق، و(1632, 768) لمجموعة الاختبار. كما تم عرض أمثلة على أسماء الأعمدة الرقمية الأولى مثل f0 إلى f9 للتحقق من سلامة ترتيب الميزات

## 4.8 تمثيل النص باستخدام Sentence-BERT (SBERT)

تم استخدام نموذج Sentence-BERT (SBERT) بإصدار all-MiniLM-L6-v2 لتحويل النصوص المستخرجة من Whisper إلى تمثيلات (Embeddings) عددية دلالية ثابتة الطول تعكس المعنى العام للجمل بدلاً من الاعتماد على الكلمات المفردة فقط. تعتمد هذه العملية على تمرير النصوص عبر محوِّل لغوي ( Language Transformer) مُدرَّب مسبقاً (Pre-trained) لإنتاج متجهات (Vectors) تمثيلية كثيفة تحتوي على معلومات سياقية ودلالية عن النص. يتم تنفيذ عملية الترميز (Encoding) على دفعات لتحسين الكفاءة، مع تطبيق توحيد (Normalization) على المتجهات (Embeddings) الناتجة لضمان اتساق القيم العددية بين جميع العينات. ينتج عن هذه المرحلة متجه (Vector) نصي بطول 384 بُعداً لكل نص، يمثّل المحتوى اللغوي للأمر الصوتي بصيغة رقمية قابلة للاستخدام في نماذج التعلم الآلي اللاحقة. وتُستخدم هذه التمثيلات النصية (Text Embeddings) لاحقاً كمدخلات مستقلة أو كجزء من نظام دمج متعدد الوسائط (Multi-Modal Fusion) مع الميزات الصوتية، مما يتيح ربط المعنى اللغوي للأوامر مع التمثيل الصوتي ضمن إطار موحّد لمعالجة النوايا وتصنيف الأوامر.

## 4.9 دمج الميزات الصوتية والنصية (Audio–Text Feature Fusion)

### 4.9.1 تقليل أبعاد الميزات الصوتية باستخدام TruncatedSVD

يتم تطبيق تقنية TruncatedSVD على الميزات الصوتية المستخرجة من Whisper (Audio Embeddings) بهدف تقليل عدد الأبعاد (Dimensionality Reduction) إلى 256 بُعداً، حيث يتم تدريب محوّل SVD على مجموعة التدريب فقط، ثم استخدامه لتحويل مجموعتي التحقق والاختبار، مما ينتج تمثيلاً صوتياً مضغوطاً (Compressed Audio Representation) يحتفظ بالمعلومات الأساسية ضمن عدد أقل من الأبعاد العددية (Feature Dimensions).

### 4.9.2 دمج الميزات الصوتية مع التمثيلات النصية

بعد الحصول على التمثيلات الصوتية المخفَّضة، يتم دمجها مع التمثيلات النصية المستخرجة مسبقاً (Text Embeddings) ضمن متجه عددي موحَّد ( Unified Feature Vector) لكل عينة عبر عملية دمج أفقي (Feature Concatenation). ينتج عن هذه العملية تمثيل مشترك يجمع المعلومات الصوتية واللغوية ( Audio–Text Fusion) ضمن بنية رقمية واحدة قابلة للاستخدام في مراحل التصنيف اللاحقة.

### 4.9.3 توحيد القيم العددية للميزات المدمجة

يتم تطبيق عملية توحيد القيم العددية (Feature Scaling) على الميزات المدمجة باستخدام StandardScaler، حيث يتم ضبط القيم استناداً إلى إحصائيات مجموعة التدريب (Training Statistics) ثم استخدام نفس التحويل على بيانات التحقق والاختبار لضمان اتساق نطاق القيم العددية (Numerical Range Consistency) عبر جميع المجموعات.

### 4.9.4 إضافة ضجيج عددي إلى بيانات التدريب

بعد توحيد الميزات، يتم إضافة مقدار محدود من الضجيج العددي ( Gaussian Noise) إلى بيانات التدريب فقط باستخدام توزيع طبيعي بانحراف معياري صغير (Noise Standard Deviation)، بهدف إدخال قدر من العشوائية أثناء عملية التعلم (Regularization Effect) دون تعديل بيانات التحقق والاختبار (Validation/Test Sets).

## 4.10 تصنيف الأوامر (Command Classification)

### 4.10.1 تدريب نموذج Logistic Regression متعدد الفئات

يتم تدريب نموذج Logistic Regression متعدد الفئات على الميزات الصوتية والنصية المدمجة باستخدام خوارزمية الحل SAGA لدعم التصنيف متعدد الفئات بنمط multinomial. يعتمد النموذج على انتظام من نوع L2 للحد من فرط التكيّف، مع ضبط معامل الانتظام C بقيمة 0.03 للتحكم بدرجة تعقيد النموذج. تم تحديد الحد الأقصى لعدد التكرارات عند 12000 لضمان استقرار عملية التدريب والوصول إلى التقارب العددي، مع تعيين قيمة العتبة لإيقاف التدريب عند تحقق شرط التقارب. كما تم تفعيل المعالجة المتوازية للاستفادة من جميع الأنوية المتاحة، وتثبيت البذرة العشوائية random_state لضمان قابلية إعادة التجربة وإنتاج نتائج متسقة عبر عمليات التشغيل المختلفة. يعتمد النموذج على التمثيل العددي الناتج عن دمج الميزات الصوتية والنصية كمدخلات رئيسية لتعلم حدود القرار بين فئات الأوامر المختلفة ضمن إطار تصنيف متعدد الفئات.

## 4.11 آلية الرفض والتحقق من صلاحية الأوامر (Rejection Mechanism)

### 4.11.1 توحيد النص وتصحيح أخطاء التفريغ الصوتي (Text Normalization and Whisper Error Correction)

يتم تطبيق سلسلة من خطوات التوحيد النصي (Text Normalization) على النص الناتج عن Whisper بهدف تقليل الضجيج اللغوي وتحسين قابلية التحليل الدلالي. تشمل هذه العملية إزالة التشكيل والترميز غير القياسي باستخدام unicodedata.normalize، وتحويل النص إلى ASCII لتجنب تأثير الأحرف الخاصة، ثم تحويل الأحرف إلى lowercase، وإزالة الرموز غير الأبجدية عبر تعبيرات منتظمة (regex). بعد ذلك، يتم تطبيق قاموس تصحيحات مخصص (WHISPER_FIXES) لمعالجة الأخطاء الشائعة في مخرجات Whisper مثل quietel أو quiter وتحويلها إلى الصيغة الصحيحة quieter، إضافة إلى تصحيح عبارات صوتية مشوّهة مثل torn it up إلى turn it up، مما يحسّن دقة تفسير النية النصية قبل تمريرها إلى مراحل اتخاذ القرار.

### 4.11.2 تحويل النص إلى نية دلالية وتحديد نطاق الأمر (Intent Mapping and In-Domain Detection)

بعد التوحيد، يتم تمرير النص إلى دالة map_text_to_intent التي تقوم بتحويل النص إلى تمثيل نية (Intent Representation) يتكون من الفعل (action)، والكائن (object)، والوسم المركّب (pair). تعتمد هذه المرحلة على قواعد لغوية مرنة (Rule-based Heuristics) لاكتشاف الأوامر المرتبطة بمجالات مثل volume، music، lights، وlanguage، مع دعم حالات التحكم الدقيقة مثل quieter أو louder وربطها مباشرة بأوامر decrease volume أو increase volume. كما يتم تصنيف بعض الحالات على أنها ضمن النطاق ولكن غامضة (in_domain_ambiguous) عندما لا تكون النية محددة بشكل كافٍ، وذلك بهدف منع الرفض المبكر للأوامر التي قد تكون صالحة بعد معالجة لاحقة. في المقابل، يتم إرجاع قيمة None عندما لا يمكن مطابقة النص مع أي نية صالحة، مما يشير إلى احتمال كون الأمر خارج نطاق النظام (out-of-domain).

### 4.11.3 آلية الرفض المعتمدة على القواعد النصية (Rule-Based Text Rejection Gate)

تمثل دالة should_reject الطبقة الأساسية لاتخاذ قرار الرفض الأولي قبل تمرير الإدخال إلى المصنف الإحصائي، حيث تقوم بتصفية الحالات غير الصالحة اعتمادًا على جودة النص الناتج عن ASR. يتم رفض الحالات التي يكون فيها النص فارغًا، أو قصيرًا جدًا (أقل من حرفين)، أو غير قابل للمعالجة بسبب فشل التفريغ الصوتي (ASR failure). كما يتم رفض الحالات التي لا يمكن ربطها بأي نية دلالية صالحة بعد تنفيذ map_text_to_intent، باعتبارها أوامر خارج نطاق النظام. في المقابل، يتم السماح للحالات المصنفة على أنها in-domain ambiguous بالمرور إلى المراحل اللاحقة دون رفض مباشر، مع تسجيل ملاحظة بأنها أوامر ضمن النطاق ولكنها غير حاسمة دلاليًا، مما يحقق توازنًا بين الصرامة والمرونة في بوابة القبول النصية.

### 4.11.4 آلية اتخاذ القرار المعتمدة على الثقة الاحتمالية (Confidence-Based Rejection using Probability Thresholds)

بعد الحصول على احتمالات التصنيف من النموذج الإحصائي، يتم تطبيق دالة decide_with_thresholds لاتخاذ قرار القبول أو الرفض بناءً على مستوى الثقة (Prediction Confidence). تعتمد هذه الآلية على اختيار أعلى احتمالين من متجه الاحتمالات (top-2 probabilities)، ثم حساب الفرق بين أعلى احتمال p1 والاحتمال الثاني p2 لتقدير هامش الفصل (Confidence Margin). يتم قبول التنبؤ

فقط إذا تجاوز أعلى احتمال عتبة محددة مسبقًا thr=0.70، وكان الفرق بين الاحتمالين أكبر من margin=0.15، مما يقلل من احتمالية اتخاذ قرارات غير مستقرة في الحالات المتقاربة احتماليًا. يتم تسجيل معلومات تفصيلية عن القرار مثل التنبؤ الأول والثاني، وقيم الاحتمالات، والفجوة بينهما، وحالة الثقة النهائية، مما يوفّر طبقة تفسيرية (Decision Interpretability) يمكن الاستفادة منها في تحليل أخطاء النموذج لاحقًا.

### 4.11.5 تكامل طبقتي الرفض النصي والإحصائي ضمن خط أنابيب القرار ( Hybrid Rejection Pipeline)

تعتمد البنية النهائية لآلية الرفض في النظام على تكامل طبقتين متتاليتين: الأولى تعتمد على قواعد نصية (Text-based Filtering) لتصفية الأخطاء المبكرة الناتجة عن ASR أو النصوص غير الصالحة، والثانية تعتمد على تقييم الثقة الاحتمالية ( Model Confidence Assessment) بعد التصنيف. يتيح هذا التصميم الهجين ( Hybrid Rejection Strategy) تقليل معدلات التنبؤ الخاطئ في الحالات خارج النطاق (Out-of-Domain Commands)، وتحسين موثوقية النظام في البيئات الواقعية، مع الحفاظ على مرونة التعامل مع الأوامر الغامضة داخل النطاق دون رفضها بشكل قاطع.

# الفصل الخامس: التجارب والنتائج والتقييم

## 5.1التجارب المنفذة(Conducted Experiments)

### 5.1.1تجربة استراتيجية تقسيم البيانات ومنع تسربها

تم في مرحلة مبكرة من المشروع اعتماد أسلوب تقسيم البيانات باستخدام Stratified Splitدون فصل العينات بناءً على معرف المتحدث(speakerId) ، مما أدى إلى احتمال وجود تسجيلات لنفس المتحدث ضمن مجموعتي التدريب والاختبار. تسبب ذلك في ظهور نتائج أداء مرتفعة بشكل غير واقعي في بعض التجارب، نتيجة استفادة النموذج من خصائص صوتية خاصة بالمتحدث بدل تعلّم أنماط عامة قابلة للتعميم.

بعد ملاحظة هذا التأثير، تم اعتماد أسلوب Group Splitالمبني على speakerIdلضمان الفصل الكامل بين المتحدثين عبر مجموعات التدريب والتحقق والاختبار، مما أدى إلى الحصول على نتائج أكثر واقعية وموثوقية، وعكس الأداء الحقيقي للنموذج في سيناريوهات عملية تتضمن متحدثين جدد.

### 5.1.2تجربة دمج التفريغ النصي الأصلي مع الميزات الصوتية (Transcription & Whisper Audio Fusion)

تم تنفيذ تجربة تعتمد على دمج النص الأصلي المرافق للبيانات (Original Transcription)مع الميزات الصوتية المستخرجة من نموذج Whisperضمن تمثيل موحّد .(Feature Fusion) حققت التجربة قيمًا مرتفعة على مؤشرات الأداء الكمية (Accuracyو Macro-F1)تجاوزت 0.99 على مجموعتي التحقق والاختبار، إلا أن التقييم العملي كشف عن أخطاء دلالية في فهم النية رغم ارتفاع النتائج الرقمية.

أمثلة على حالات الخطأ:

- *Turn off the music* → Pred: change language → English

  True: deactivate → music
- *Switch off the washroom lights* → Pred: activate → music

  True: deactivate → lights

تشير هذه النتائج إلى وجود تضارب دلالي ناتج عن دمج نص من مصدر مختلف عن مصدر الصوت، إضافةً إلى تأثير أخطاء التفريغ وعدم الاتساق بين التمثيلين النصي والصوتي. كما يُحتمل أن تكون النتائج الرقمية المرتفعة في هذه المرحلة قد تأثرت جزئيًا باستراتيجية التقسيم المبكرة، إلا أن قرار استبعاد هذا النهج في النموذج النهائي جاء بناءً على تحليل أخطاء التنبؤ العملية وليس المؤشرات الرقمية فقط.

### 5.1.3تجربة دمج ميزات Whisper الصوتية مع النص الناتج عنه باستخدامTF-IDF

اعتمدت هذه التجربة على دمج الميزات الصوتية المستخرجة من Whisperمع النص الناتج عن التفريغ الصوتي نفسه، بعد تمثيله باستخدام TF-IDFلاستخراج سمات نصية إحصائية قبل تنفيذ التصنيف. أظهرت النتائج تحسنًا ملحوظًا في دقة التنبؤ مقارنةً بالتجربة السابقة، خصوصًا في الأوامر المباشرة والواضحة.

أمثلة على تنبؤات صحيحة:

- *Decrease volume* → decrease → volume (Confidence ≈ 0.999)
- *Play Music, please!* → activate → music (Confidence ≈ 0.999)
- *Switch the keyboard to English* → change language → English

في المقابل، استمر النموذج في إظهار أخطاء عند التعامل مع الأوامر غير المباشرة أو الغامضة مثل:

- *Turn it up* → deactivate → music (خطأ دلالي)
- *Louder please* → activate → music بدلincrease → volume

تشير هذه النتائج إلى أن TF-IDFيعزز الأداء في الأوامر الصريحة، لكنه يظل محدودًا في فهم النية الضمنية أو الصياغات غير المباشرة.

### 5.1.4تجربة استخدام ميزات MFCC و Delta و Delta-Delta قبلWhisper

تم تنفيذ تجربة تعتمد على استخراج ميزات صوتية تقليدية من الإشارة الخام باستخدام MFCCإلى جانب Deltaو .Delta-Deltaأظهرت نتائج العرض باستخدام PCAوجود تداخل كبير بين الفئات المختلفة وعدم تحقق فصل واضح بين النوايا، مما يدل على أن هذه الميزات تلتقط الخصائص الفيزيائية العامة للصوت، لكنها غير كافية لتمثيل المعنى الدلالي أو التمييز بين النوايا المتقاربة لغويًا.

### 5.1.5آلية الرفض الأساسية(Baseline Rejection Mechanism)

تعتمد آلية الرفض الأساسية على تحليل النص الناتج عن التفريغ الصوتي باستخدام قواعد لغوية بسيطة تهدف إلى تحديد ما إذا كان الأمر ضمن نطاق الأوامر المدعومة أم خارج المجال. تقوم الآلية بتوحيد النص، ثم مطابقة كلمات مفتاحية تمثل نوايا محددة. في حال عدم وجود مؤشرات دلالية كافية، يتم رفض الأمر تلقائيًا.

تتميز هذه المقاربة بالبساطة وسهولة التنفيذ، لكنها محدودة في التعامل مع أخطاء التفريغ، الصياغات غير المباشرة، أو الأوامر الغامضة، مما قد يؤدي إلى قرارات رفض أو قبول غير دقيقة.

### 5.2 معايير تقييم أداء النموذج(Evaluation Metrics)

تم تقييم أداء النموذج باستخدام مجموعة من المقاييس الكمية المعيارية المناسبة لمهام تصنيف النوايا الصوتية. شملت هذه المقاييس **الدقة (Accuracy)** لقياس نسبة التنبؤات الصحيحة إجمالًا، و**درجة F1 الماكرو (Macro F1)** لقياس التوازن بين الدقة والاسترجاع عبر جميع الفئات بشكل متساوٍ، بما يحدّ من تأثير عدم توازن البيانات بين الفئات.
كما تم الاعتماد على **تقرير التصنيف (Classification Report)** لتحليل قيم Precision و Recallو F1لكل فئة على حدة، إضافةً إلى **مصفوفة الالتباس (Confusion Matrix)** لدراسة أنماط الأخطاء وحالات الالتباس بين النوايا المتقاربة دلاليًا.
ولتقييم سلوك النموذج مع تغيّر حجم بيانات التدريب، تم استخدام **منحنيات التعلم (Learning Curves)**التي توضح تطور الأداء على مجموعات التدريب والتحقق والاختبار مع زيادة عدد العينات.

### 5.3 تقييم أداء النموذج باستخدام المقاييس الكمية(Quantitative Model Evaluation)

أظهرت نتائج التقييم الكمي للنموذج النهائي المعتمد (SBERT + Whisper Audio + Logistic Regression)أداءً مرتفعًا على مجموعة الاختبار، حيث بلغت الدقة الكلية (Accuracy)قيمة 97.12%على عدد إجمالي قدره 1632عينة اختبار .كما حقق النموذج Macro F1 = 96.55%، مما يشير إلى توازن جيد في الأداء عبر جميع فئات النوايا دون انحياز ملحوظ للفئات الأكثر تمثيلًا.

على مستوى الفئات الفردية، حقق النموذج أداءً مرتفعًا في معظم الأوامر الصوتية، إذ بلغت قيمة F1-scoreلفئة change language → English حوالي 99.28%مع Precision = 99.28% و Recall = 99.28%، وهو أعلى أداء بين جميع الفئات.
كما سجّلت فئة activate → lights قيمة F1-score = 97.50%، بينما حققت فئة deactivate → lights قيمة 97.36%، مما يدل على قدرة النموذج على التمييز الجيد بين أوامر الإضاءة المتقابلة.

في المقابل، انخفض الأداء نسبيًا في بعض الفئات ذات الدعم العددي الأقل أو القريبة دلاليًا، مثل فئة activate → music التي حققت Precision = 89.47% و F1-score = 92.90%،

وهو ما يعكس وجود بعض حالات الخلط الدلالي بينها وبين أوامر أخرى مرتبطة بالموسيقى أو مستوى الصوت.
كما حققت فئتا decrease → volume وincrease → volume قيم F1-score بلغت 96.81% و96.40% على التوالي، مع توازن جيد بين قيم Precision و Recall.

يوضح تقرير التصنيف (Classification Report) أن قيم Recall كانت مرتفعة نسبيًا في معظم الفئات، حيث تجاوزت 96% في جميع الأوامر تقريبًا، مما يدل على قدرة النموذج على اكتشاف النوايا الصحيحة حتى في وجود ضوضاء أو تنوع لغوي.
كما تؤكد مصفوفة الالتباس (Confusion Matrix) أن الأخطاء المتبقية تتركز بشكل أساسي بين النوايا المتقاربة دلاليًا، مثل أوامر التحكم بالموسيقى ومستوى الصوت، وليست ناتجة عن أخطاء عشوائية في التصنيف.

بشكل عام، تشير هذه النتائج إلى أن النموذج يحقق أداءً كميًا قويًا ومتوازنًا، مع فجوة محدودة بين الفئات، إلا أن وجود بعض الانخفاضات في Precision لبعض الأوامر يؤكد الحاجة إلى دعم النتائج الرقمية بتحليل عملي ونوعي للأداء في سيناريوهات الاستخدام الواقعي.

.

### 5.3 التقييم العملي للنظام باستخدام بيانات واقعية(Real-World Evaluation)

تم تقييم النظام باستخدام مجموعة تسجيلات صوتية واقعية مستقلة عن بيانات التدريب، تم تسجيلها في بيئات متنوعة تتضمن ضوضاء خلفية، موسيقى، وحديث جانبي. شملت مجموعة التقييم أوامر ضمن النطاق(Accept) ، وأوامر خارج النطاق(Reject) ، وأوامر غامضة أو غير مباشرة.(Ambiguous In-Domain)

أظهرت نتائج التنفيذ النهائي قدرة النظام على تنفيذ الأوامر المباشرة بدقة وثقة مرتفعتين، إضافةً إلى نجاحه في تفسير بعض الأوامر غير المباشرة مثل:

- *No more lights* → deactivate → lights
- *Stop* → deactivate → music

في المقابل، استمرت بعض الإخفاقات عند التعامل مع تعبيرات اصطلاحية أو صياغات غير قياسية مثل:

- *Crank it up* → activate → music بدل increase → volume

كما نجحت آلية الرفض المحسّنة في استبعاد الأوامر غير المرتبطة بالمجال، مما قلّل من خطر تنفيذ قرارات خاطئة في الاستخدام الواقعي.

## 5.4 تحليل الفجوة بين النتائج الرقمية والأداء الواقعي(Performance Gap Analysis)

على الرغم من ارتفاع النتائج الرقمية، أظهر التقييم العملي وجود فجوة محدودة لكنها مهمة بين الأداء الإحصائي والسلوك الواقعي للنظام. تعود هذه الفجوة إلى اختلاف طبيعة البيانات الواقعية التي تتضمن تنوعًا لغويًا، وضوضاء صوتية، وتعبيرات غير مباشرة، مقارنةً ببيانات التدريب المنظمة.

كما أن بعض الأخطاء المتبقية كانت دلالية بطبيعتها، حيث ينجح النموذج في التقاط الفئة العامة لكنه يخطئ في تمييز النية الدقيقة عند تقارب المعاني بين الأوامر، مما يؤكد ضرورة الجمع بين التقييم الكمي والتحليل النوعي.

## 5.5 تأثير طبيعة البيانات وبنية المهمة على الأداء

ساهم التنظيم الجيد للبيانات وعدد النوايا المحدود (سبع فئات رئيسية فقط) في تحقيق أداء مرتفع جدًا على بيانات التدريب، حيث اقتربت بعض المقاييس من القيمة الكاملة. إلا أن نتائج التحقق والاختبار أظهرت فجوة محدودة وغير مقلقة، مما يشير إلى أن النموذج لا يعتمد على الحفظ السلبي بل يتعلم أنماطًا قابلة للتعميم.

كما أكدت نتائج التقييم العملي قدرة النظام على التعامل مع بيانات غير مرئية في بيئات متنوعة، مع بقاء تحديات في الحالات غير المباشرة أو المتقاربة دلاليًا.

## 5.6 خلاصة التجارب العامة(Overall Experimental Findings Summary)

أظهرت التجارب أن أفضل أداء تحقق عند اعتماد ميزات صوتية عميقة مستخرجة من Whisperمع النص الناتج عنه، ثم تمثيل النص دلاليًا باستخدام SBERT، يلي ذلك تصنيف النية باستخدام Logistic Regressionكنموذج فعّال وخفيف.

ساهم استخدام TruncatedSVDفي تقليل أبعاد الميزات الصوتية وتحسين الاستقرار وتقليل الضجيج العددي. كما لعبت تقنيات Regularizationدورًا مهمًا في تقليل فرط التخصيص وتحسين قدرة النموذج على التعميم.

أثبت دمج الميزات الصوتية والنصية من نفس المصدر فعاليته مقارنةً باستخدام ميزات تقليدية مثل MFCC، كما حسّنت آلية الرفض المحسّنة من موثوقية النظام عبر استبعاد الأوامر خارج النطاق ومعالجة الحالات الغامضة بأمان أكبر.

بشكل عام، حقق النظام توازنًا عمليًا بين الأداء الرقمي المرتفع والدقة الدلالية الواقعية، مع استمرار بعض التحديات المرتبطة بالتعبيرات غير المباشرة وتقارب النوايا.

## 5.7 التحديات التي واجهت عملية التدريب(Training Challenges)

شملت التحديات الرئيسية عدم توازن البيانات بين الفئات، تفاوت جودة التسجيلات الصوتية، وجود ضوضاء بيئية، تشابه النوايا دلاليًا بين بعض الأوامر، صعوبة التعامل مع الصياغات غير المباشرة، إضافةً إلى مشكلة تسرب البيانات في مرحلة مبكرة بسبب تقسيم غير معتمد على speakerId.

## 5.8 التحديات التي تم التغلب عليها(Challenges Mitigated)

تمت معالجة هذه التحديات عبر اعتماد Group Split لمنع تسرب البيانات، استخدام TruncatedSVD لتحسين تمثيل الميزات، تطبيق Regularization للحد من فرط التخصيص، تعزيز التمثيل النصي باستخدام SBERT، وتطوير آلية رفض محسّنة لتحسين موثوقية النظام في السيناريوهات الواقعية.

# الفصل السادس:الخاتمة و الافاق المستقبلية:

## 6.1 الخاتمة

بعد تنفيذ سلسلة التجارب وتطوير نظام التعرف على الأوامر الصوتية، تبيّن أن الجمع بين ميزات Whisper الصوتية والنص المستخرج باستخدام Whisper ASR، مع تمثيل دلالي عبر SBERTوتصنيف باستخدام Logistic Regression، قد حقق أداءً قويًا من الناحية الكمية، مع قدرة ملحوظة على التعميم على عينات غير مرئية سابقًا.

أظهرت النتائج أن النموذج قادر على التمييز بين معظم الأوامر الأساسية بدقة مرتفعة، مع فجوة محدودة بين أداء التدريب والتحقق والاختبار، ما يشير إلى أن النموذج لا يعتمد فقط على حفظ البيانات، بل يتعلّم أنماطًا عامة قابلة للتعميم. كما ساهم استخدام تقليل الأبعاد باستخدام SVD وتنظيم النموذج (Regularization) في تحسين الاستقرار وتقليل فرط التخصيص.

مع ذلك، كشفت التجارب العملية عن وجود أخطاء دلالية متبقية، خصوصًا في الحالات التي تتضمن أوامر متقاربة في المعنى أو صياغات غير مباشرة، بالإضافة إلى تحديات في التعامل مع الضجيج البيئي وأخطاء التفريغ الصوتي. لذلك، يمكن اعتبار النظام ناجحًا كنموذج أولي عملي، مع وجود هامش واضح للتحسين في سيناريوهات الاستخدام الواقعي.

بشكل عام، يثبت المشروع إمكانية بناء نظام تحكم صوتي فعّال باستخدام نماذج حديثة لمعالجة الصوت والنص، مع التأكيد على أهمية الجمع بين التقييم الرقمي والتقييم العملي للوصول إلى صورة دقيقة عن الأداء الحقيقي للنظام.

## 6.2 التوصيات المستقبلية

- زيادة عدد الأوامر المدعومة لتغطية نطاق أوسع من السيناريوهات الواقعية، مما يقلل من حالات الغموض والتداخل الدلالي بين النوايا.
- دمج نماذج لغوية كبيرة (LLMs) في طبقة الفهم الدلالي، بهدف تحسين تفسير الأوامر غير المباشرة، وتعزيز القدرة على فهم الصياغات المتنوعة والسياق اللغوي، خاصة في الحالات التي تفشل فيها القواعد أو التمثيلات الدلالية المحدودة.

- تعزيز آليات التنظيم (Regularization) مثل زيادة معامل الانتظام، واستخدام تقنيات إضافية مثل Dropout وWeight Decay، وذلك لتقليل احتمالية فرط التخصيص وتحسين استقرار الأداء على بيانات غير مرئية.
- تحسين آلية رفض الأوامر خارج النطاق (Out-of-Domain Rejection) عبر استخدام أساليب أكثر تقدمًا لقياس عدم اليقين (Uncertainty Estimation)، مما يقلل من الأخطاء الناتجة عن التنبؤات الواثقة غير الصحيحة.
- تدريب النموذج على بيانات تحتوي مستويات ضجيج أعلى وأكثر تنوعًا، لتحسين الأداء في البيئات الواقعية الصاخبة.
- تحسين أداء النظام في الزمن الحقيقي من خلال تقليل زمن الاستجابة وتحسين كفاءة استهلاك الموارد، تمهيدًا لنشر النظام كتطبيق عملي.

## 5.3 الرؤية المستقبلية للنظام

يمثّل هذا المشروع خطوة أولى نحو تطوير أنظمة تحكم صوتي ذكية ومحددة النطاق يمكن استخدامها في تطبيقات عملية مثل:

- التحكم في إعدادات الأجهزة الذكية
- المساعدات الصوتية المخصصة لنطاق محدود
- أنظمة دعم المستخدم في التطبيقات البرمجية
- واجهات تحكم صوتي في البيئات التعليمية أو المكتبية

وبفضل مرونة البنية المعتمدة وإمكانية توسيعها مستقبلًا، يمكن تطوير النظام ليصبح منتجًا تطبيقيًا قابلًا للنشر بعد تحسين التغطية الدلالية للأوامر وتعزيز الأداء في الظروف الواقعية المتنوعة.

## المراجع:

[1] Das, N., Dingliwal, S., Ronanki, S., et al. (2024). SpeechVerse: A Large-scale Generalizable Audio-Language Model. arXiv.

[2] Everson, K., Gu, Y., Yang, H., et al. (2024). Towards ASR-Robust Spoken Language Understanding Using Word Confusion Networks . arXiv.

[3]Benayas, A. (2024). Enhancing Intent Classifier Training with Large Language Model-generated Data . International Journal of Human-Computer Studies.

[4] .Pekarek Rosin, T., Kaplan, B. C., & Wermter, S. (2025). LLM Data Generation for Intent Recognition in German Speech .arXiv.

[5] Zhu, Z., Zhang, F., Sun, J., et al. (2025). Novel Utterance Data Augmentation via Large Language Models. Springer.

[6] Zhang, Y., Wu, Q., Chen, X., et al. (2025). A Survey on Multi-modal Intent Recognition. Findings of EMNLP 2025.

[7] Huang, R., Li, T., & Chen, B. (2025). Multi-modal Intent


**Syrian private university**

**Faculty of Engineering**

**Artificial intelligence and data Science**

**Applications**

# Voice Command Recognition

Prepared by :

Raneem Rabih Ayham Alsalem

Supervisors:

Dr. Maissa Aboukassem Eng. Wessam Alsohli