

أعدت هذه الأطروحة

لإنجاز مقرر المشروع الفصلي في اختصاص الذكاء الصناعي 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 الكمبيوتر 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 المستخدم

## 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

7	قائمة بأهم المصطلحات :
10	قائمة بالاشكال :
12	الفصل الأول:
12	1.1 مقدمة عن المشروع :
12	1.2 الهدف من المشروع :
12	1.3 المشكلة التي يقوم المشروع بحلها
14	الفصل الثاني :
14	2.1 الدراسة النظرية:
14	Whisper
15	Sentence-BERT (SBERT)
16	Logistic Regression
16	TruncatedSVD (Dimensionality Reduction)
17	TF-IDF (Text Representation)
17	Rejection Mechanism (Decision Layer)
17	الفصل الثالث: الدراسة المرجعية
17	3.1 مقدمة
18	3.2 الدراسة الأولى – SpeechVerse :نموذج صوت-لغة واسع النطاق
18	3.3 الدراسة الثانية Word Confusion Networks :لتحسين متانة SLU
18	3.5 الدراسة الرابعة: توليد البيانات باستخدام LLM لتصنيف النية باللغة الألمانية
18	3.6 الدراسة الخامسة: توليد عبارات جديدة باستخدام LLM لتوسيع بيانات SLU
19	3.8 الدراسة السادسة: فهم النية متعدد الوسائط (EMNLP 2025)
21	الفصل الرابع: منهجية العمل
21	4.1 مقدمة
21	4.2 منهجية العمل العامة
22	4.3 مجموعة البيانات المستخدمة
22	4.4 تحضير البيانات (Data Preparation)

22	4.4.1 حذف الأعمدة غير الضرورية من مجموعات البيانات
23	4.4.2 إنشاء المسارات الكاملة للملفات الصوتية
23	4.4.3 تحويل النصوص المفرّغة إلى أوامر معيارية
23	4.4.4 تحديث حقول الأوامر استناداً إلى نتائج المعالجة النصية
23	4.4.5 تصفية الصفوف المرتبطة بكائنات غير مستهدفة
24	4.4.6 إنشاء تسمية مركبة للأوامر وعرض توزيعها الإحصائي
	4.4.7 تقسيم البيانات باستخدام Group Split بناءً على المتحدث ( Speaker-Based Splitting)
24	4.4.8 تحميل بيانات التدريب المخصّصة ومعالجة الـ Labels
24	4.4.9 تعريف دالة تعزيز الإشارة الصوتية (Audio Augmentation)
25	4.4.10 توليد عينات إضافية للوصول إلى عدد مستهدف لكل فئة
25	4.4.11 دمج العينات الأصلية مع العينات المعززة
25	4.4.12 تقليل العينات الزائدة للوصول إلى حجم موحد لكل فئة
25	4.4.13 عرض التوزيع الإحصائي للفئات قبل وبعد الموازنة
26	4.5 استخراج الميزات الصوتية (Feature Extraction)
26	4.5.1 تحميل نموذج Whisper والمعالج المرافق له
26	4.5.2 تحميل الملفات الصوتية وتحويلها إلى صيغة Mono بمعدل 16kHz
26	4.5.3 تنفيذ التجميع المتوسط مع مراعاة القناع الزمني (Masked Mean Pooling)
27	4.5.4 استخراج ميزات Whisper باستخدام إعدادات إدخال محددة
27	4.5.5 تجميع المتجهات الصوتية وتحويلها إلى تمثيل ثابت الطول
27	4.5.6 حفظ الميزات المستخرجة وربطها بالبيانات التعريفية
27	4.5.7 تنفيذ عملية الاستخراج على مجموعات التدريب والتحقق والاختبار
28	4.6 استخراج النص باستخدام Whisper (Whisper ASR Transcription)
28	4.6.1 تجهيز مجلدات التخزين المؤقت والحفظ النهائي
28	4.6.2 تحميل ملفات الميزات والتحقق من أبعاد التقسيمات
28	4.6.3 تحميل نموذج Whisper وتنفيذ التفريغ الصوتي
29	4.6.4 إضافة النصوص المفرّغة وإدارة الكاش أثناء التنفيذ
29	4.6.5 أبعاد البيانات بعد إضافة whisper_text
29	4.6.6 حفظ ملفات الميزات المحدثات
29	4.7 تحضير البيانات ومواءمة الميزات (Data Preparation and Feature Alignment)

29	4.7.1 تحميل الميزات الصوتية والتحقق من الأبعاد
30	4.7.2 توحيد النصوص والـ Labels وتجهيز الحقول الأساسية
30	4.7.4 دمج الميزات الصوتية وبناء مجموعات البيانات النهائية
30	4.7.5 إعداد الـ Labels النهائية وقوائم النصوص لكل تقسيم
31	4.7.6 توحيد وترتيب الأعمدة الرقمية للميزات الصوتية
31	4.7.7 إنشاء مصفوفات الميزات الصوتية بعد المواءمة
31	4.8 تمثيل النص باستخدام Sentence-BERT (SBERT)
32	4.9 دمج الميزات الصوتية والنصية (Audio-Text Feature Fusion)
32	4.9.1 تقليل أبعاد الميزات الصوتية باستخدام TruncatedSVD
32	4.9.2 دمج الميزات الصوتية مع التمثيلات النصية
32	4.9.3 توحيد القيم العددية للميزات المدمجة
32	4.9.4 إضافة ضجيج عددي إلى بيانات التدريب
33	4.10 تصنيف الأوامر (Command Classification)
33	4.10.1 تدريب نموذج Logistic Regression متعدد الفئات
33	4.11 آلية الرفض والتحقق من صلاحية الأوامر (Rejection Mechanism)
	4.11.1 تطبيع النص وتصحيح أخطاء التفريغ الصوتي (Text Normalization and Whisper)
33	(Error Correction)
	4.11.2 تحويل النص إلى نية دلالية وتحديد نطاق الأمر (Intent Mapping and In-)
34	(Domain Detection)
34	4.11.3 آلية الرفض المعتمدة على القواعد النصية (Rule-Based Text Rejection Gate)
	4.11.4 آلية اتخاذ القرار المعتمدة على الثقة الاحتمالية (Confidence-Based Rejection)
35	(using Probability Thresholds)
	4.11.5 تكامل طبقتي الرفض النصي والإحصائي ضمن خط أنابيب القرار (Hybrid)
35	(Rejection Pipeline)
36	الفصل الخامس: التجارب والنتائج والتقييم
36	5.1 التجارب المنفذة (Conducted Experiments)
36	5.1.1 تجربة استراتيجية تقسيم البيانات ومنع تسرب البيانات
	5.1.2 تجربة دمج التفريغ النصي الأصلي مع الميزات الصوتية (Transcription & Whisper)
36	Audio Fusion)
37	5.1.3 تجربة دمج ميزات Whisper الصوتية مع النص الناتج عنه باستخدام TF-IDF
37	5.1.4 تجربة استخدام ميزات MFCC و Delta و Delta-Delta قبل Whisper

37	آلية الرفض الأساسية (Baseline Rejection Mechanism)
38	تقييم أداء النموذج باستخدام المقاييس الكمية (Quantitative Model Evaluation)
38	التقييم العملي للنظام باستخدام بيانات واقعية (Real-World Evaluation)
39	تحليل الفجوة بين النتائج الرقمية والأداء الواقعي (Performance Gap Analysis)
39	تأثير طبيعة البيانات وبنية المهمة على الأداء
39	5.6 خلاصة التجارب العامة (Overall Experimental Findings Summary)
40	5.7 التحديات التي واجهت التدريب (Training Challenges)
40	5.8 التحديات التي تم التغلب عليها (Challenges Mitigated)
41	الفصل السادس: الخاتمة و الافاق المستقبلية:
41	6.1 الخاتمة
41	6.2 التوصيات المستقبلية
42	5.3 الرؤية المستقبلية للنظام
43	المراجع:

## قائمة بأهم المصطلحات :

المصطلح التقني	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 وفعال
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	مميزات 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 احتماليين 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	لكل F1 و Recall و Precision تقرير يعرض فئة
Accuracy	نسبة التنبؤات الصحيحة إجمالاً
Precision	نسبة التنبؤات الصحيحة داخل فئة معينة
Recall	نسبة الحالات الصحيحة التي تم اكتشافها
F1 Score	Precision و Recall مقياس يوازن بين
Macro F1	عبر جميع الفئات بالتساوي F1 متوسط
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	Whisper أخطاء ناتجة عن تفريغ
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

## قائمة بالاشكال :

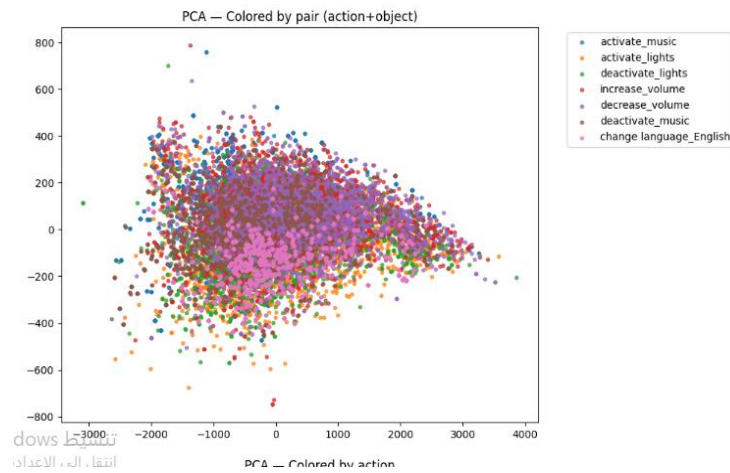


Figure 1

	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

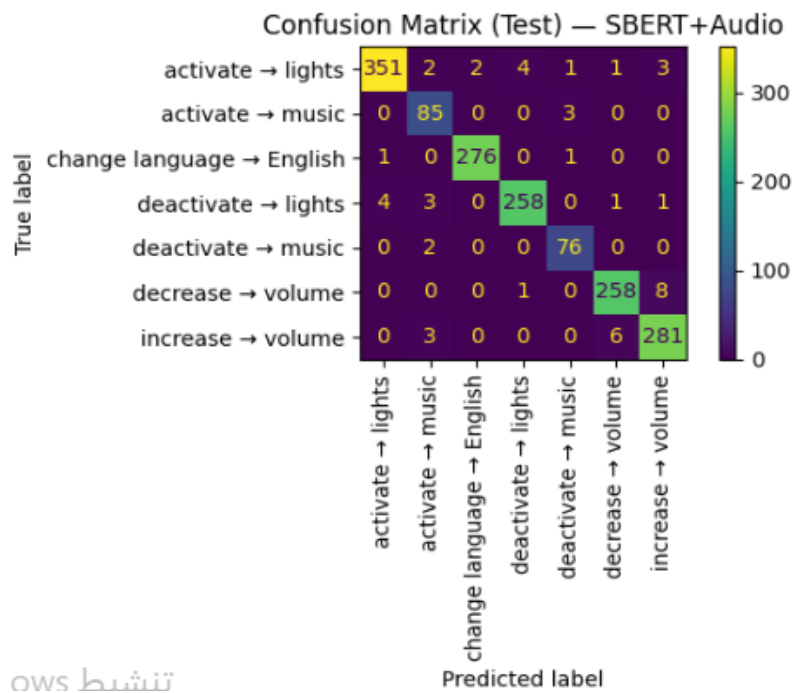


Figure 3

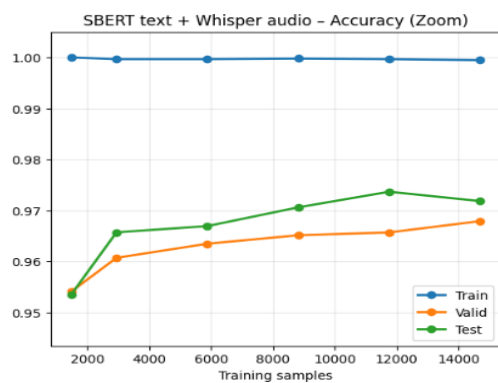


Figure5

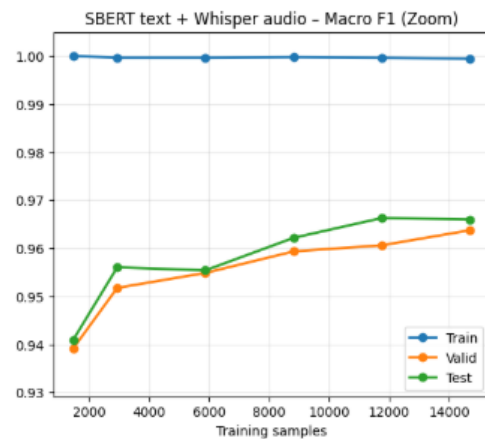


Figure4

## الفصل الأول:

### 1.1 مقدمة 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. يركز المشروع على فهم الأمر الصوتي وظيفيًا وليس فقط تحويل الصوت إلى نص، مما يجعله أكثر ملائمة 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 ( $Action \rightarrow Object$ ) بدل الاكتفاء بالنص الخام.

### 4. بناء نموذج Fusion يجمع بين:

◦ الميزات الصوتية المستخرجة من Whisper

◦ الميزات النصية المستخرجة من نص Whisper

### 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 Reasoning لاحقًا 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 الكمبيوتر 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 اليدين مشغولتين، يصبح التحكم الصوتي وسيلة مثالية للتفاعل مع الحاسوب، إذ يوفر للمستخدم مرونة أكبر ويجعل تجربة الاستخدام أكثر سلاسة وفعالية.

## الفصل الثاني :

### 2.1 الدراسة النظرية:

#### Whisper

يُعد نموذج Whisper من النماذج المتقدمة في مجال التعرف التلقائي على الكلام (Automatic Speech Recognition – ASR)، وهو نموذج عميق قائم على بنية Transformer Encoder–Decoder تم 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 Encoder الذي 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 Decoder بتوليد النص المقابل خطوة بخطوة من خلال آلية الانتباه المتقاطع (Cross-Attention) التي تربط التمثيلات الصوتية المخرجة من الـ Encoder بالسلسلة النصية الناتجة، بحيث يتم التنبؤ بالكلمة التالية اعتماداً على السياق الصوتي والسياق اللغوي السابق في آن 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 التوقعات، واستنتاج الكلمات غير الواضحة اعتماداً على السياق العام للجملة. إضافة إلى إنتاج النص، يستطيع Whisper استخراج تمثيلات عديدة عميقة (Audio Embeddings) تمثل جوهر الإشارة الصوتية بشكل مضغوط، وهي متجهات رقمية تحتوي معلومات مركبة 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 تعلمي مباشر من البيانات، مما يجعلها أكثر قدرة على تمثيل المعنى الصوتي الفعلي بدلاً من الاكتفاء بالخصائص الفيزيائية الخام. وبذلك، يشكّل Whisper نظاماً متكاملًا 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 مُشَقَّر (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 للجمل 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 عددياً ثابت الأبعاد يعكس العلاقات النحوية، المعنى السياقي، والبنية الدلالية العامة للجمل، 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 تعبر 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 للتعلم، ثم تمرير الناتج عبر دالة لوجستية لتحويله إلى قيمة احتمالية محصورة بين 0 و 1. 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 متوقع.

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 التكيف عبر تقليل تأثير الأوزان الكبيرة، مما يساعد على تحسين قدرة النموذج على التعميم على بيانات جديدة. يتم التحكم بدرجة التنظيم من خلال المعامل  $C$ ، 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 تمثيلًا للمعلومات الأساسية. تعتمد هذه الطريقة على تفكيك المصفوفة الأصلية إلى عدد محدود من المركبات الخطية (Singular Vectors) التي تفسر 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 التعلم 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 فعال 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 الدراسة الأولى – SpeechVerse: نموذج صوت-لغة واسع النطاق

تناولت الدراسة الأولى تطوير نموذج متكامل يجمع بين معالجة الإشارة الصوتية والنمذجة اللغوية العميقة ضمن إطار موحد، أطلق عليه اسم **SpeechVerse**. يهدف هذا النموذج إلى فهم الكلام مباشرة دون الاعتماد على مراحل منفصلة للتعرف على الصوت ثم تحليل النص.

اعتمد الباحثون على مجموعات بيانات صوتية كبيرة ومتنوعة، شملت لهجات وبيئات مختلفة، وتم تدريب النموذج على مهام متعددة مثل التعرف على الكلام، فهم النية، والتفاعل مع المستخدم. أظهرت نتائج الدراسة تحسناً ملحوظاً في الدقة تراوح بين 9% و 23% مقارنة بالنماذج التقليدية، خاصة في معالجة الأوامر القصيرة والمتشابهة صوتياً.

### 3.3 الدراسة الثانية Word Confusion Networks: لتحسين متانة SLU

ركزت الدراسة الثانية على مشكلة أخطاء أنظمة التعرف التلقائي على الكلام (ASR) وتأثيرها السلبي على فهم النية. اقترح الباحثون استخدام شبكات التباس الكلمات (**Word Confusion Networks**) كتمثيل وسيط يدمج عدة فرضيات ناتجة عن ASR بدل الاعتماد على نص واحد فقط.

تم تقييم النموذج على بيانات صوتية متنوعة تحتوي على مستويات عالية من الضوضاء، وأظهرت النتائج تحسناً واضحاً في مقاومة الأخطاء مقارنة بالأساليب التقليدية. تعكس هذه الدراسة أهمية التعامل مع عدم يقين مخرجات Whisper أو ASR.

### 3.5 الدراسة الرابعة: توليد البيانات باستخدام LLM لتصنيف النية باللغة الألمانية

استعرضت الدراسة الرابعة دور النماذج اللغوية الكبيرة في تحسين تصنيف النية للغات محدودة الموارد، مع التركيز على اللغة الألمانية. اعتمد الباحثون على توليد أمثلة لغوية جديدة تغطي تراكيب وأساليب مختلفة، ثم تدريب نماذج تصنيف تقليدية على البيانات الأصلية والموسعة.

أظهرت النتائج تفوقاً واضحاً للنماذج المدربة على البيانات الموسعة، مما يبرز قدرة LLM على معالجة مشكلة محدودية الموارد اللغوية.

### 3.6 الدراسة الخامسة: توليد عبارات جديدة باستخدام LLM لتوسيع بيانات SLU

قدمت الدراسة الخامسة منهجية جديدة تعتمد على توليد عبارات لغوية متنوعة باستخدام LLM بهدف زيادة تنوع بيانات فهم اللغة المنطوقة. تم تطبيق المنهجية على مجموعات بيانات صغيرة الحجم، وأظهرت النماذج المدربة على البيانات الموسعة أداءً أفضل في التعامل مع العبارات غير المرئية سابقاً.

### 3.8 الدراسة السادسة: فهم النية متعدد الوسائط (EMNLP 2025)

استعرضت الدراسة السابعة أحدث الاتجاهات في نماذج فهم النية متعددة الوسائط، مع التركيز على دمج الصوت والنص والسياق الدلالي ضمن نموذج موحد. أظهرت النتائج أن النماذج المعتمدة على LLM متعددة الوسائط تحقق أعلى مستويات الدقة، خاصة في السيناريوهات المعقدة التي تتطلب فهماً عميقاً للسياق.

النتائج الرئيسية	النموذج الأفضل	Dataset	المنهج أو التقنية المستخدمة	معايير التقييم	عنوان الورقة	الرقم
أداء مرتفع في تعميم النوايا الصوتية عبر مجالات متعددة	SpeechVerse Large	SpeechVerse Dataset	Audio Transformer + LLM Fusion	Intent Accuracy, Zero-shot	SpeechVerse: A Large-scale Generalizable Audio-Language Model (2024)	1
تحسين مقاومة أخطاء التفريغ الصوتي بشكل ملحوظ	WCN + GPT-based LLM	SLURP, FSC	ASR → WCN → LLM	Robust Accuracy, WER	ASR-Robust SLU via Word Confusion Networks (2024)	2
رفع الدقة وتحسين التعميم عبر البيانات الاصطناعية	RoBERTa + LLM Augmentation	Banking77, CLINC150	LLM Data Generation + BERT / RoBERTa	Intent Accuracy, F1-Score	Enhancing Intent Classifier using LLM-generated Data (2024)	3
تحسن الأداء عند استخدام بيانات مولدة	LLM-Generated Data Pipeline	German SLU Datasets	Whisper Fine-Tuned + LLM	Intent F1, ASR Robustness	LLM Data Generation for German Intent Recognition (2025)	4
تقليل الالتباس بين النوايا وتحسين الفصل الدلالي	Transformer + Paraphrase-LLM	SNIPS, ATIS	GPT-style Paraphrasing + Transformers	Accuracy, Intra-class Variance	Novel Utterance Data Augmentation via LLMs (2025)	5
أداء قوي عند دمج النص والصوت في تمثيل موحد	MIR Fusion Transformer	MIR-2025	Audio Encoder + LLM + Cross-modal Layers	Accuracy, Cross-modal Fusion Gain	Multi-modal Intent Recognition (MIR) – EMNLP 2025	6

## الفصل الرابع: منهجية العمل

### 4.1 مقدمة

تُعد أنظمة التعرف على الأوامر الصوتية من أكثر تطبيقات الذكاء الاصطناعي تحديًا، 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 (FSC) Fluent Speech Commands من مجموعات البيانات المرجعية في مجال فهم الأوامر الصوتية (Spoken Language Understanding)، حيث تحتوي على تسجيلات صوتية لـ 97 متحدثًا قاموا بنطق 248 عبارة مختلفة تمثل 31 نية (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 المقيمين، وتم استبعاد أي ملفات تحتوي على ضجيج مرتفع، أو صوت غير واضح، أو أخطاء في العبارة المنطوقة. يتم تنظيم البيانات ضمن مجلدات بحسب معرف المتحدث (Speaker ID)، بحيث يحتوي كل مجلد على ملفات صوتية بصيغة WAV بمعدل أخذ عينات 16 كيلوهرتز وقناة واحدة، إضافة إلى ملفات توصيف بصيغة CSV لكل من مجموعات التدريب والتحقق والاختبار، تتضمن معلومات تفصيلية لكل عينة مثل مسار الملف الصوتي، معرف المتحدث، النص المنطوق، والفئات الدلالية المرتبطة به، حيث تشمل قيم Action أوامر مثل (change language, activate, deactivate, increase, decrease, bring)، وقيم Object عناصر مثل (music, lights, volume, lamp, newspaper, juice, socks, shoes, English, German)، 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

يقوم هذا الجزء من الكود بحذف العمود location من مجموعات بيانات التدريب والتحقق والاختبار باستخدام الدالة drop من مكتبة Pandas، مع تفعيل الخيار errors='ignore' لضمان عدم حدوث أخطاء في حال عدم توفر العمود. كما يتم حذف العمود Unnamed: 0 عند وجوده، وهو عمود يظهر عادةً نتيجة تصدير البيانات بصيغة CSV ويتضمن فهرس الصفوف. بعد تنفيذ هذه العمليات، تبقى في

البيانات الأعمدة الأساسية المرتبطة بالمحتوى الفعلي دون الأعمدة الإضافية غير اللازمة

#### 4.4.2 إنشاء المسارات الكاملة للملفات الصوتية

يتم تعريف دالة باسم `make_full` تقوم بربط المسار النسبي لكل ملف صوتي مع المسار الأساسي `base_path` باستخدام `os.path.join`. بعد ذلك، يتم إنشاء عمود جديد داخل إطار البيانات `df` باسم `full_path` يحتوي على المسارات الكاملة للملفات الصوتية، مما يتيح استخدام هذه المسارات للوصول إلى الملفات أثناء مراحل المعالجة اللاحقة. في النهاية، يتم عرض عينة من القيم المخزنة في هذا العمود للتحقق من صحة عملية تكوين المسارات

#### 4.4.3 تحويل النصوص المفرّغة إلى أوامر معيارية

تُعرّف دالة `map_command` لمعالجة النصوص الموجودة في العمود `transcription`، حيث يتم تحويل النص إلى أحرف صغيرة لتوحيد الصيغة النصية، ثم يتم فحص محتواه لاكتشاف كلمات مفتاحية مرتبطة بمجالات مختلفة مثل الإضاءة، تغيير اللغة، التحكم بالصوت، أو تشغيل الموسيقى. بناءً على الكلمات المكتشفة، تقوم الدالة بإرجاع زوج من القيم يمثلان الفعل (`action`) والكائن (`object`) بما يتوافق مع معنى النص، وفي حال عدم تطابق النص مع أي من القواعد المحددة، يتم الإبقاء على القيم الأصلية المسجلة في الصف دون تعديل

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

يتم تطبيق الدالة `map_command` على جميع الصفوف داخل إطار البيانات باستخدام الدالة `apply`، حيث يتم استخراج القيم الناتجة عن معالجة النص وإسنادها إلى العمودين `action` و `object`. يؤدي ذلك إلى تحديث تمثيل الأوامر داخل مجموعة البيانات بناءً على النصوص المفرّغة مع الحفاظ على هيكل إطار البيانات الأصلي دون تغيير

#### 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 و object في سلسلة نصية واحدة تفصل بينهما علامة سهم (→)، بحيث يمثل هذا العمود الصيغة الموحدة للأمر. بعد ذلك، يتم استخدام الدالة value\_counts لعرض عدد التكرارات لكل قيمة داخل هذا العمود بهدف إظهار توزيع الأوامر ضمن مجموعة البيانات

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

يستخدم هذا الجزء من الكود الصنف 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. بعد ذلك، يُنشأ عمود pair في حال لم يكن موجوداً مسبقاً، حيث يتم دمج قيم action و object في تمثيل نصي موحد. 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. في النهاية، يتم عرض أحجام كل مجموعة وعدد المتحدثين المشتركين بينها للتحقق من أن عملية الفصل تمت دون تداخل

#### 4.4.8 تحميل بيانات التدريب المخصصة ومعالجة الـ Labels

يتم تحديد مجلد يحتوي على ملفات التقسيم المحفوظة مسبقاً، ثم تحميل ملف التدريب train\_group.csv باستخدام مكتبة Pandas. بعد تحميل البيانات، يتم إنشاء عمود pair عبر دمج قيم action و object 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)

يتم تعريف الدالة augment\_audio لمعالجة الإشارة الصوتية المحملة من الملفات الصوتية، حيث تبدأ العملية بتطبيع سعة الإشارة لمنع التشبع الرقمي، ثم يتم تطبيق مجموعة من التحويلات العشوائية على الإشارة مثل الإزاحة الزمنية عبر الإزاحة الزمنية للإشارة Temporal Shift، وتغيير سرعة التشغيل باستخدام time\_stretch، وتعديل طبقة الصوت عبر pitch\_shift، وإضافة ضجيج عشوائي منخفض الشدة.



تُنفَّذ هذه العمليات باحتمالات مختلفة لكل تحويل، وتُرجع الدالة الإشارة الصوتية المعدلة لاستخدامها لاحقاً في إنشاء عينات جديدة

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

يتم تجميع بيانات التدريب وفق العمود pair، ثم حساب عدد العينات للفئة المستهدفة TARGET المحددة مسبقاً. في حال كان عدد العينات في فئة معينة أقل من العدد المطلوب، يتم تحديد عدد العينات الناقصة وسحب عينات من نفس الفئة مع إمكانية التكرار عند الحاجة. لكل عينة يتم اختيارها، يتم تحميل الملف الصوتي المرتبط بها، ثم تمرير الإشارة الصوتية إلى دالة augment\_audio لتوليد نسخة معدلة. بعد ذلك، يتم حفظ الملف الصوتي الجديد داخل مجلد AUG\_DIR مع اسم جديد مشتق من اسم الفئة والملف الأصلي، ثم يتم إنشاء صف بيانات جديد يشير إلى المسار الجديد ويتم تمييزه على أنه عينة معززة عبر تعيين قيمة العمود is\_aug إلى واحد.

#### 4.4.11 دمج العينات الأصلية مع العينات المعززة

يتم إنشاء إطار بيانات جديد aug\_df يحتوي على جميع الصفوف الناتجة عن عملية التعزيز، ثم يتم دمجها مع إطار البيانات الأصلي train\_df باستخدام pd.concat لتكوين مجموعة بيانات موسعة تضم العينات الأصلية والعينات المعززة معاً. بعد الدمج، تصبح جميع العينات متاحة ضمن إطار موحد يمكن استخدامه في عمليات التصنيف أو إعادة التوازن اللاحقة.

#### 4.4.12 تقليل العينات الزائدة للوصول إلى حجم موحد 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 موحد وإعادة ترتيب الصفوف عشوائياً لضمان توزيع متوازن.

#### 4.4.13 عرض التوزيع الإحصائي للفئات قبل وبعد الموازنة

يتم استخدام مكتبة Matplotlib لرسم مخططين عموديين يوضحان توزيع عدد العينات لكل فئة ضمن العمود pair قبل تنفيذ عملية الموازنة وبعدها. في كل مخطط، يتم تعيين عناوين واضحة للمحاور وتسمية الفئات بشكل مناسب، مع تدوير تسميات الفئات على المحور الأفقي (Rotating Category Labels) لتحسين قابلية القراءة ومنع تداخل النصوص عند عرض عدد كبير من الفئات. بعد ذلك، يتم عرض

الرسومات لمقارنة التوزيع العددي للفئات عبر مراحل تحضير البيانات المختلفة وتحليل تأثير عملية الموازنة على تقليل عدم التوازن بين الفئات.

## 4.5 استخراج الميزات الصوتية (Feature Extraction)

### 4.5.1 تحميل نموذج Whisper والمعالج المرافق له

يتم تحميل نموذج Whisper من مستودع 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 16kHz

تُعرّف دالة لتحميل الملفات الصوتية باستخدام مكتبة Librosa، حيث يتم تحويل كل ملف إلى قناة واحدة (mono=True) وتوحيد معدل أخذ العينات عند 16,000 هرتز (sr=16000)، ثم تحويل الإشارة الناتجة إلى Tensor بصيغة float32 لاستخدامها في إطار PyTorch. بعد تنفيذ هذه العملية على مجموعات البيانات، بلغت أبعاد التمثيل الصوتي الخام (773, 14700) لمجموعة التدريب، و(773, 1806) لمجموعة التحقق، و(773, 1632) 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)

يتم تعريف دالة masked\_mean\_pool لحساب المتوسط الزمني لمتجهات الحالة الخفية الناتجة عن مشفر Whisper، مع استخدام قناع الانتباه (attention\_mask) عند توفره لتجاهل القيم الناتجة عن الحشو (padding). تعتمد العملية على ضرب المتجهات بالقناع، ثم حساب مجموع القيم عبر البعد الزمني وقسمتها على عدد الإطارات الفعلية بعد تطبيق حد أدنى ثابت (clamp(min=1e-6)) لضمان الاستقرار العددي. في حال عدم توفر القناع أو عدم تطابق الأبعاد بين القناع ومخرجات المشفر، يتم استخدام المتوسط الزمني التقليدي كخيار بديل (fallback).

#### 4.5.4 استخراج ميزات Whisper باستخدام إعدادات إدخال محددة

يتم تعريف دالة رئيسية لاستخراج الميزات الصوتية لكل مجموعة بيانات، حيث يتم تحميل الإشارة الصوتية، ثم تمريرها إلى WhisperProcessor باستخدام إعدادات تشمل `sampling_rate=16000` و `padding=True` وإرجاع المخرجات بصيغة Tensor. يتم استخراج مصفوفة الخصائص الطيفية `input_features` ذات الشكل `(1, 80, T)`، مع تطبيق حشو إضافي في حال كان طول الإشارة أقل من الحد الأدنى المحدد `MIN_T=3000` لضمان ثبات طول الإدخال بين العينات. بعد ذلك، يتم تمرير الخصائص إلى مشفر Whisper باستخدام `model.get_encoder()` مع تعطيل حساب التدرجات (`torch.no_grad()`) لاستخراج المتجهات الخفية الأخيرة (`last_hidden_state`) ذات الشكل `(1, 768, T_enc)`.

#### 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.

#### 4.5.6 حفظ الميزات المستخرجة وربطها بالبيانات التعريفية

يتم تحويل المتجهات المستخرجة إلى إطار بيانات (DataFrame) يحتوي على الأعمدة العددية للميزات الصوتية، إلى جانب الأعمدة التعريفية المرتبطة بكل عينة مثل `action`، `object`، `pair`، ومسار الملف الصوتي `full_path`. كما يتم حفظ معرف المتحدث `speakerId` ومؤشر العينات المعززة `is_aug` عند توفرهما لدعم التحليل والتشخيص اللاحق. بعد ذلك، يتم حفظ الملفات الناتجة بصيغة CSV داخل مجلد الإخراج المحدد لكل تقسيم باستخدام تسمية `whisper_train.csv` و `whisper_valid.csv` و `whisper_test.csv`.

#### 4.5.7 تنفيذ عملية الاستخراج على مجموعات التدريب والتحقق والاختبار

في المرحلة النهائية، يتم تنفيذ دالة استخراج الميزات على مجموعات البيانات الثلاث (`train`، `valid`، `test`)، حيث يتم عرض أبعاد كل مجموعة قبل بدء المعالجة، ثم تتبع تقدم التنفيذ باستخدام `tqdm` مع تسجيل عدد الملفات المفقودة التي لم يُعثر على

مساراتها، وعدد الملفات القصيرة التي تم تطبيق الحشو عليها. بعد انتهاء التنفيذ، يتم حفظ المسارات النهائية للملفات الناتجة وعرض ملخص بعدد الميزات المستخرجة لكل تقسيم.

## 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 الميتاداتا بعد إضافة النصوص (تمثل ملفات الـ `Metadata` في هذا المشروع جداول وصفية تحتوي على النصوص المستخرجة من نموذج `Whisper` إلى جانب الـ `Labels` المرتبطة بكل عينة، والمسارات الصوتية، والمعلومات التعريفية الأخرى. وتستخدم 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 الميتاداتا والتحقق من أبعاد التقسيمات

تم تحميل ملفات الميتاداتا 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` و `object` في تمثيل نصي موحد لضمان اتساق الـ `Labels` 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 و 50 ثانية لمعالجة 14700 ملف صوتي، بمعدل يقارب 4.55 ملفات في الثانية. كما

استغرقت عملية تفريغ مجموعة التحقق حوالي 8 دقائق و 20 ثانية لمعالجة 1806 ملفات، بينما استغرقت مجموعة الاختبار حوالي 7 دقائق و 24 ثانية لمعالجة 1632 ملفاً صوتياً.

#### 4.6.4 إضافة النصوص المفرّغة وإدارة الكاش 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 الكاش، حيث بلغت قيمة العناصر المسترجعة من الكاش 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 و 10 أعمدة بعد إضافة عمود النص المفرّغ، في حين أصبحت مجموعة التحقق تحتوي على 1806 صفوف و 9 أعمدة، ومجموعة الاختبار تحتوي على 1632 صفاً و 9 أعمدة. تم التأكد من نجاح إضافة العمود الجديد إلى جميع التقسيمات قبل الانتقال إلى مراحل تمثيل النصوص واستخراج الميزات الدلالية.

#### 4.6.6 حفظ ملفات الميئاتا المحدثّة

تم حفظ ملفات الميئاتا بعد إضافة النصوص المفرّغة بصيغة CSV داخل مجلد الحفظ النهائي، بحيث تم إنشاء ملفات مستقلة لكل من مجموعة التدريب والتحقق والاختبار تحتوي على النصوص الناتجة عن Whisper، ثم تم عرض المسار النهائي للمجلد لتأكيد مكان تخزين الملفات الناتجة التي ستُستخدم لاحقاً في مراحل تمثيل النصوص النصية واستخراج الميزات الدلالية.

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

#### 4.7.1 تحميل الميئاتا والميزات الصوتية والتحقق من الأبعاد

تم تحميل ملفات الميئاتا الخاصة بمجموعات التدريب والتحقق والاختبار، حيث بلغت الأبعاد (14700, 10) لمجموعة التدريب، و(1806, 9) لمجموعة التحقق،

و(9, 1632) لمجموعة الاختبار بعد تضمين النصوص الناتجة عن Whisper وال Labels 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 وال Labels وتجهيز الحقول الأساسية

تم التأكد من توفر العمود whisper\_text ضمن ملفات الميزات، مع معالجة القيم المفقودة وتحويلها إلى صيغة نصية موحدة، كما تم توحيد أعمدة action و object 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 مفرغ 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 الميزات 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 الميزات الصوتية وبناء مجموعات البيانات النهائية

تم دمج الأعمدة النصية وال Labels من ملفات الميزات الصوتية 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، ال Labels، والميزات العددية ضمن بنية موحدة (Unified Dataset Structure).

#### 4.7.5 إعداد ال Labels النهائية وقوائم النصوص لكل تقسيم

تم استخراج ال Labels 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 (768, 14700) لمجموعة التدريب، و(768, 1806) لمجموعة التحقق، و(768, 1632) 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 محوّل 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 تطبيع (Normalization) على المتجهات (Embeddings) الناتجة لضمان اتساق القيم العددية بين جميع العينات. ينتج عن هذه المرحلة متجه (Vector) نصي بطول 384 بُعداً لكل نص، يمثل المحتوى اللغوي للأمر الصوتي بصيغة رقمية قابلة للاستخدام في نماذج التعلم الآلي اللاحقة. وتُستخدم هذه التمثيلات النصية (Text Embeddings) لاحقاً كمداخلات 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 موحد 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 المخفّضة، يتم دمجها مع التمثيلات النصية المستخرجة مسبقاً (Text Embeddings) ضمن متجه عددي موحد (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 (`solver="saga"`) لدعم التصنيف متعدد الفئات بنمط `multinomial`. يعتمد النموذج على انتظام من نوع `L2` (`penalty="l2"`) للحد من فرط التكيف، مع ضبط معامل الانتظام `C` بقيمة `0.03` للتحكم بدرجة تعقيد النموذج. تم تحديد الحد الأقصى لعدد التكرارات عند `12000` (`max_iter=12000`) لضمان استقرار عملية التدريب والوصول إلى التقارب العددي، مع تعيين قيمة العتبة `tol=1e-3` لإيقاف التدريب عند تحقق شرط التقارب. كما تم تفعيل المعالجة المتوازية عبر `n_jobs=-1` 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 تطبيع النص وتصحيح أخطاء التفريغ الصوتي (Text Normalization and Whisper Error Correction)

يتم تطبيق سلسلة من خطوات التطبيع 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 quieter أو quiter وتحويلها إلى الصيغة الصحيحة quieter، إضافة إلى تصحيح عبارات صوتية مشوهة مثل torn it up إلى turn it up، مما يحسن 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 التطبيق، 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 المركب (pair). تعتمد هذه المرحلة على قواعد لغوية مرنة (Rule-based Heuristics) لاكتشاف الأوامر المرتبطة بمجالات مثل volume، music، lights، language، 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 بتصنيف 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  $p_1$  والاحتمال الثاني  $p_2$  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 يوفر 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

### 5.1 التجارب المنفذة (Conducted Experiments)

#### 5.1.1 تجربة استراتيجية تقسيم البيانات ومنع تسرب البيانات

تم في مرحلة مبكرة من المشروع اعتماد أسلوب تقسيم البيانات باستخدام Stratified Split دون 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 الاستفادة النموذج من خصائص صوتية خاصة بالمتحدث بدل تعلم أنماط عامة قابلة للتعميم.

بعد ملاحظة هذا التأثير، تم اعتماد أسلوب Group Split المبني على speakerId لضمان 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.

#### 5.1.2 تجربة 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 Whisper ضمن تمثيل موحد (Feature Fusion). حققت التجربة قيمًا مرتفعة على مؤشرات الأداء الكمية (Accuracy) و (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، إضافة 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.

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

اعتمدت هذه التجربة على دمج الميزات الصوتية المستخرجة من Whisper مع النص الناتج عن التفريغ الصوتي نفسه، بعد تمثيله باستخدام TF-IDF لاستخراج سمات نصية إحصائية قبل تنفيذ التصنيف. أظهرت النتائج تحسناً ملحوظاً في دقة التنبؤ مقارنةً بالتجربة السابقة، خصوصاً 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 بدل increase → volume

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

### 5.1.4 تجربة استخدام ميزات MFCC و Delta و Delta-Delta قبل Whisper

تم تنفيذ تجربة تعتمد على استخراج ميزات صوتية تقليدية من الإشارة الخام باستخدام MFCC إلى جانب Delta و Delta-Delta. أظهرت نتائج التصوير باستخدام PCA وجود 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 لغوياً.

### 5.1.5 آلية 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 بتطبيع النص، ثم مطابقة كلمات مفتاحية تمثل نوايا محددة. في حال عدم وجود مؤشرات دلالية كافية، يتم رفض الأمر تلقائياً.

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 تقييم أداء النموذج باستخدام المقاييس الكمية (Quantitative Model Evaluation)

تم تقييم النموذج النهائي المعتمد (SBERT + Whisper Audio + Logistic Regression) باستخدام مقاييس معيارية تشمل Accuracy و Macro F1 على مجموعات التدريب والتحقق والاختبار. أظهرت النتائج أداءً مرتفعاً على مجموعة التدريب (Train Accuracy  $\approx 99.95\%$ ، Train Macro-F1  $\approx 99.95\%$ ، مما يشير إلى قدرة النموذج على تمثيل بيانات التدريب بشكل شبه كامل.

في المقابل، انخفض الأداء على مجموعة التحقق (Validation Accuracy  $\approx 96.73\%$ ، Validation Macro-F1  $\approx 96.32\%$ ) واستقر عند مستوى قريب على مجموعة الاختبار (Test Accuracy  $\approx 97.12\%$ ، Test Macro-F1  $\approx 96.55\%$ )، مما يعكس فجوة أداء محدودة لكنها ملموسة بين التدريب والتعميم الفعلي.

يوضح Classification Report أن معظم الفئات تحقق قيم Precision و Recall مرتفعة، بينما تُظهر Confusion Matrix استمرار بعض حالات الالتباس بين نوايا متقاربة دلاليًا مثل التحكم بالموسيقى ومستوى الصوت، مما يؤكد أن الأخطاء المتبقية ليست عشوائية بل مرتبطة بتقارب المعنى. كما تُظهر Learning Curves أن زيادة عدد العينات حسّنت الاستقرار دون تحقيق تطابق كامل بين التدريب والاختبار.

## 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 الديمو 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:

• increase → volume بدل Crank it up → activate → music

كما نجحت آلية الرفض المحسنة في استبعاد الأوامر غير المرتبطة بالمجال، مما قلل 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 تنوعاً لغوياً، 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 جداً 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 أنماطاً 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 Whisper مع النص الناتج عنه، ثم تمثيل النص دلاليًا باستخدام SBERT، يلي ذلك تصنيف النية باستخدام Logistic Regression كنموذج فعال وخفيف.

ساهم استخدام TruncatedSVD في تقليل أبعاد الميزات الصوتية وتحسين الاستقرار وتقليل الضجيج العددي. كما لعبت تقنيات Regularization دورًا 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 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.



### 6.1 الخاتمة

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

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 يتعلم أنماطاً 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، خصوصاً 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 ناجحاً كنموذج أولي عملي، مع وجود هامش واضح للتحسين في سيناريوهات الاستخدام الواقعي.

بشكل عام، يثبت المشروع إمكانية بناء نظام تحكم صوتي فعال 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 و Weight 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

يمثل 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 مستقبلاً، يمكن تطوير النظام ليصبح منتجاً تطبيقياً قابلاً للنشر بعد تحسين التغطية الدلالية للأوامر وتعزيز الأداء في الظروف الواقعية المتنوعة.

- [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