الجامعة السورية الخاصة
كلية هندسة الذكاء الصنعي
قسم هندسة الذكاء الصنعي وعلوم البيانات الذكية

نظام دبلجة آلي للفيديوهات باستخدام تقنيات التعلم العميق
(من الإنجليزية إلى العربية)

Automated Dubbing System for Videos using Deep Learning Techniques (English to Arabic)

أعدت هذه الأطروحة

لإنجاز مشروع التخرج في اختصاص الذكاء الصنعي وعلوم البيانات

اعداد الطلاب :

حمزة زاهر السمان

رغد محمد علي عبد الرحمن

أسماء المشرفين :

د.ماجدة البكور

م.وسام السحلي

العام الدراسي 2026/2025

الملخص :

يتناول هذا المشروع تصميم وتنفيذ نظام دبلجة آلي متكامل للفيديوهات التعليمية، يهدف إلى **نقل المحتوى الصوتي والمرئي** من اللغة الإنجليزية إلى العربية باستخدام تقنيات التعلم العميق. تبرز المسألة البحثية في فجوة التواصل اللغوي التي تحد من الوصول للمحتوى الأكاديمي العالمي، والحاجة لنظم دبلجة سريعة وواقعية تتجاوز عقبات التكلفة والزمن في الطرق التقليدية.

تكمن الإضافة النوعية لهذا البحث في بناء وتدريب نموذج شبكة عصبونية عميقة تعتمد على بنية الـ U-Net المتخصصة في فصل الضجيج وتعزيز الكلام(Speech Enhancement) ؛ حيث تم تطوير البنية وتدريبها لضمان استخلاص إشارة صوتية نقية من الفيديوهات التعليمية قبل البدء بالمعالجة اللغوية، وهو ما يميز هذا العمل عن الدراسات المرجعية التي تعتمد غالباً على نماذج جاهزة.

يتكامل هذا النموذج المدرب مع pipeline متطور يعتمد على تقنيات WhisperX لتحقيق التعرف الآلي على الكلام والترجمة. وقد تم توسيع قدرات النظام ليشمل تقنية **فصل المتحدثين (Speaker Diarization)** وتحديد جنسهم بدقة، مما يسمح بتخصيص الأصوات المُولدة لتناسب كل متحدث، مع الاعتماد على المحاذاة الزمنية الدقيقة (Word-level Alignment) لضمان مطابقة الصوت المترجم مع الجدول الزمني للفيديو الأصلي.

ولارتقاء جودة العمل إلى مستوى احترافي متكامل بَصرياً وصوتياً، تمت إضافة مرحلة **مزامنة الشفاه (Lip-syncing)** بالاعتماد على معمارية .MuseTalk ولتجاوز تحديات التشوه البصري، تم ابتكار خوارزمية تنعيم رياضية

(Smoothing Logic) لمعالجة اهتزاز إحداثيات الوجه (Jittering) ، وتطبيق نموذج **CodeFormer** المتقدم لترميم وتحسين جودة ملامح الوجه بدقة فائقة.

أظهرت النتائج العملية كفاءة عالية للنظام في تحسين جودة الصوت، وتقليل الأخطاء الزمنية، وتوليد حركة شفاه متطابقة تماماً مع النص العربي. حيث أثبتت التجارب قدرة المنظومة على معالجة ودبلجة فيديو تعليمي مدته 8 دقائق في زمن 12 دقيقة باستخدام معالج رسوميGPU ، مما يجعله حلاً فعالاً وقابلاً للتوسع في **دبلجة منصات التعليم المفتوح وإنتاج محتوى ناطق بالعربية بجودة عالية**.

Abstract :

This project presents the design and implementation of an integrated automated **dubbing** system for educational videos, aimed at transferring both audio and visual content from English to Arabic using deep learning techniques. The research addresses the linguistic communication gap that limits access to global academic content, highlighting the need for fast and realistic dubbing systems that overcome the cost and time constraints of traditional methods.

The core contribution of this work lies in building and training a custom deep neural network based on the **U-Net** architecture specifically for speech enhancement and noise separation. By developing this architecture, the system ensures the extraction of clean audio signals from educational videos before linguistic processing, distinguishing this work from existing studies that often rely on pre-built models.

This model is integrated into an advanced pipeline utilizing **WhisperX** for Automatic Speech Recognition (ASR) and translation. The system's capabilities have been extended to include **Speaker Diarization** and precise **Gender Detection** utilizing Wav2Vec2-based classification, allowing for the assignment of context-appropriate synthetic voices to each speaker. Furthermore, **Word-level Alignment** is employed to ensure the translated audio perfectly matches the original video timeline.

To achieve professional-grade audiovisual output, a **Lip-syncing** phase was implemented using the **MuseTalk** architecture. To overcome visual distortions, a novel mathematical **Smoothing Logic** was developed to stabilize facial coordinates and eliminate jittering. Additionally, the advanced **CodeFormer** model was applied for high-fidelity facial restoration and feature enhancement, ensuring a seamless and high-quality visual experience.

Experimental results demonstrate high system efficiency in improving audio quality, reducing temporal errors, and generating lip movements that are perfectly synchronized with the Arabic text. Tests confirm the system's ability to process and dub an 8-minute educational video in approximately 12 minutes using GPU acceleration, providing an effective and scalable solution for dubbing open education platforms and producing high-quality Arabic-spoken content.

الفهرس

الملخص : 2
فهرس المصطلحات 6
الفصل الاول - المقدمة 11
1.1تمهيد 11
1.2مشكلة البحث 11
1.3أهداف البحث 11
1.4 الجهات المستفيدة 11
1.5النتائج والإسهامات 12
1.6هيكلية البحث 12
الفصل الثاني –الأسس التقنية للدبلجة والأدبيات السابقة 14
2.1 المفاهيم الأساسية 14
2.2 تطور الدبلجة من التقليدية إلى الذكاء الاصطناعي 14
2.3 علاقة الذكاء الاصطناعي بمجال الدبلجة 15
2.4 الأدبيات السابقة 15
الفصل الثالث – التقنيات والأدوات المستخدمة 19
3.1 المفاهيم التقنية والخصائص الأساسية: 19
3.1.1تقنية Demucs لفصل المصادر الصوتية: 19
3.1.2 نموذج ResUNet المطور (النموذج الخاص بنا): 19
3.1.3تقنية WhisperX للمحاذاة والتعرف على الكلام: 19
3.1.4المقارنة بين Whisper (الأصلي) و WhisperX (المطور) : 19
3.1.5تقنية Pyannote.audio لفصل المتحدثين: 20
3.1.6تقنية Wav2Vec2 لتصنيف جنس المتحدث: 20
3.1.7تقنيات الترجمة اللغوية (LLM) والمعالجة النصية: 20
3.1.8تقنيات توليد النطق والاستنساخ : 21
3.1.9معمارية MuseTalk للمزامنة البصرية: 21
3.1.10معمارية CodeFormer لترميم وتحسين معالم الوجه: 21
الفصل الرابع– منهجية البحث والمنظومة المقترحة 23
4.1مجموعة المعطيات المستخدمة وتحضيرها 23
4.2 مخطط المنهجية العام: 24
4.3 التحليل الهندسي لبنية نموذج ResUnet المطور 25

4.4 التحسينات التي قدمها هذا المنهج مقارنة بالأساليب التقليدية: 27
الفصل الخامس– التجارب والنتائج والتقييم 29
5.1التحديات والاحتياجات: 29
5.2 آلية عمل الاختبار والتدريب: 29
5.3مقاييس التقييم: 30
5.4نتائج الاختبار وتحليل الأداء: 30
الفصل السادس -الخاتمة والآفاق المستقبلية..... 33
6.1 الخاتمة: 33
6.2 الآفاق المستقبلية : 33
6.3التوصيات: 34
6.4 الملاحظات الختامية: 34
References المراجع..... 35

فهرس المصطلحات

المصطلح التقني (بالإنجليزية)	الترجمة للعربية	الاختصار	المعنى
Deep Learning	التعلم العميق	**DL**	فرع من الذكاء الاصطناعي يعتمد على شبكات عصبية معقدة لمحاكاة التعلم البشري.
Convolutional Neural Network	الشبكات التلافيفية	**CNN**	نوع من الشبكات العصبية العميقة المصممة لمعالجة البيانات الشبكية (مثل المخططات الطيفية للصوت).
U-Net	شبكة يو-نت	**U-Net**	بنية شبكة عصبونية عميقة تعتمد على المشفّر وفك التشفير (Encoder-Decoder) وتستخدم بكفاءة في فصل الإشارات.
Automatic Speech Recognition	التعرف الآلي على الكلام	**ASR**	عملية تحويل الكلام المنطوق من إشارة صوتية إلى نص مكتوب.
Speech-Noise Separation	فصل الكلام عن الضجيج	**-**	تقنية تهدف لعزل صوت المتحدث الأساسي عن ضجيج الخلفية لتحسين جودة الصوت.
WhisperX	ويسبر-إكس	**-**	نموذج متطور مبني على Whisper من OpenAI، مخصص للتعرف على الكلام والترجمة مع محاذاة زمنية دقيقة.
Short-Time Fourier Transform	تحويل فوريه قصير المدى	**STFT**	عملية رياضية تستخدم لتحويل الإشارة الصوتية من المجال الزمني إلى المجال الترددي (المخطط الطيفي).
Signal-to-Noise Ratio	نسبة الإشارة إلى الضجيج	**SNR**	مقياس يستخدم لتقييم جودة الصوت عبر مقارنة قوة الإشارة المرغوبة بمستوى الضجيج.
Voice Activity Detection	كشف نشاط الصوت	**VAD**	تقنية تستخدم لتحديد الأجزاء التي تحتوي على نطق بشري في التسجيل الصوتي واستبعاد فترات الصمت.
Temporal Alignment	المحاذاة الزمنية	**-**	عملية مطابقة النص المترجم مع الطوابع الزمنية الدقيقة لمقطع الفيديو الأصلي لضمان تزامن الدبلجة.

المصطلح التقني (بالإنجليزية)	الترجمة للعربية	الاختصار	المعنى
Word Error Rate	نسبة الخطأ في الكلمة	**WER**	المقياس الأساسي لتقييم دقة أنظمة التعرف على الكلام (ASR)، حيث يقيس الفرق بين النص الناتج والنص الأصلي1
Mean Squared Error	متوسط مربع الخطأ	**MSE**	دالة خسارة تستخدم غالباً في تدريب شبكات U-Net لقياس الفرق بين الصوت النقي والصوت الناتج عن النموذج.
Loss Function	دالة الخسارة	**-**	دالة رياضية تقيس مدى دقة تنبؤات النموذج أثناء عملية التدريب لتقليل الخطأ تدريجيا.
Optimizer	المحسّن	**-**	خوارزمية (مثل Adam) تُستخدم لتحديث أوزان الشبكة العصبية لتقليل دالة الخسارة أثناء التدريب
Inference Time	زمن الاستدلال	**-**	الوقت الذي يستغرقه النموذج لمعالجة البيانات (مثل الفيديو) وإخراج النتيجة النهائية.
Encoder-Decoder	المشفّر وفك التشفير	**-**	بنية هندسية تعتمد عليها شبكة U-Net؛ حيث يقوم المشفّر بضغط البيانات وفك التشفير بإعادة بنائها
Skip Connections	وصلات التخطي	**-**	ميزة في بنية U-Net تسمح بنقل المعلومات مباشرة بين الطبقات المتقابلة للحفاظ على التفاصيل الدقيقة للصوت
Validation Set	مجموعة التحقق	**-**	جزء من البيانات يُستخدم لاختبار النموذج أثناء التدريب للتأكد من قدرته على التعميم ومنع فرط التخصيص

فهرس الأشكال

الشكل **الصفحة**

الشكل (4-1): المخطط الانسيابي العام لمنهجية عمل نظام الدبلجة الآلي **24**

الشكل (4-2): بنية ResUnet **26**

الشكل (5-1): منحنيات تطور دالة الخسارة ومقياس SI-SDR أثناء تدريب النموذج **30**

الشكل (5-2): القيم الإحصائية التي حققها النظام عبر مقاييس الجودة العالمية **31**

فهرس الجداول

الجداول	الصفحة
جدول (1-2) : مقارنة تقنية بين منهجيات ومعايير قياس الأبحاث السابقة	16
جدول (1-3) : مقارنة بين whisper & whisperx	20
جدول (1-4) : إحصائيات تقسيم مجموعة البيانات المستخدمة	23
جدول (1-5) : مقاييس تقييم الأداء المستخدمة في المشروع	30

الفصل الأول

المقدمة

الفصل الاول - المقدمة

1.1 تمهيد

يشهد العصر الحالي تحولاً رقمياً هائلاً في مجال التعليم المفتوح، حيث أصبحت المنصات التعليمية العالمية مصدراً أساسياً للمعرفة. ومع ذلك، تظل اللغة عائقاً رئيسياً يحول دون وصول شريحة واسعة من المتعلمين الناطقين بالعربية إلى هذا المحتوى التعليمي المتقدم. يهدف مشروع "نظام الدبلجة الآلي" إلى توظيف تقنيات التعلم العميق ومعالجة الإشارات الصوتية لسد هذه الفجوة المعرفية؛ وذلك عبر نقل المحتوى الصوتي والمرئي للفيديوهات التعليمية من الإنجليزية إلى العربية آلياً، مع ضمان دقة الترجمة، وملاءمة الأصوات، وتطابق حركة الشفاه، لجعل التعلم المفتوح متاحاً للجمهور العربي بكفاءة وواقعية .

1.2 مشكلة البحث

تواجه أنظمة دبلجة الفيديوهات التقليدية والآلية عدة تحديات تعيق اعتمادها بشكل واسع، أبرزها تأثر جودة التعرف على الكلام بوجود الضجيج والتشويش في التسجيلات الأصلية، بالإضافة إلى صعوبة الفصل بين المتحدثين وتخصيص أصوات مناسبة لكل جنس لضمان واقعية الدبلجة .كما تبرز معضلة التزامن اللفظي والبصري، حيث يصعب مطابقة النص المترجم مع الطوابع الزمنية للفيديو وحركة شفاه المتحدث بشكل دقيق .فضلاً عن مشكلة التكلفة والزمن في الدبلجة البشرية. يعالج هذا البحث هذه المشكلات عبر بناء منظومة متكاملة تبدأ بنموذج U-Net مُدرب خصيصاً لفصل الضجيج، متبوعاً بنظام WhisperXلضمان الترجمة وفصل المتحدثين، وصولاً إلى استخدام تقنيات MuseTalk و CodeFormer لتحقيق مزامنة شفاه احترافية وترميم ملامح الوجه بدقة عالية

1.3 أهداف البحث

يسعى هذا البحث لتحقيق مجموعة من الأهداف التقنية والتطبيقية، وهي:

1. تطوير نموذج ذكاء اصطناعي قادر على تنقية الكلام البشري من الضجيج البيئي والموسيقى الخلفية لرفع دقة المعالجة/..

2. أتمتة عملية الترجمة والمحاذاة الزمنية وفصل المتحدثين وتحديد جنسهم آلياً لضمان مطابقة الصوت المترجم مع هوية المتحدث والجداول الزمنية للفيديو .

3. تحقيق التوافق البصري الكامل عبر مزامنة حركة الشفاه باستخدام معمارية MuseTalk وابتكار خوارزميات لتنعيم حركة الوجه ومنع الاهتزاز .

4. رفع الجودة البصرية للمخرجات باستخدام نموذج CodeFormer لترميم ملامح الوجه وضمان دبلجة سينمائية احترافية .

5. بناء نظام ذو كفاءة زمنية عالية يتيح معالجة الفيديوهات التعليمية الطويلة في زمن قياسي وتقليل الفجوة اللغوية.

1.4 الجهات المستفيدة

يستهدف هذا المشروع شريحة واسعة من القطاعات والأفراد في العالم العربي، حيث يساهم بشكل مباشر في كسر الحواجز اللغوية وتسهيل الوصول إلى المعرفة الأكاديمية العالمية. وتتلخص الجهات المستفيدة من هذا النظام فيما يلي:

- **الطلاب والأكاديميون العرب :** يتيح النظام للطلاب في مختلف المراحل الدراسية (لا سيما طلاب الهندسة والعلوم المتقدمة) الوصول إلى المحاضرات والمساقات العالمية المتاحة على منصات التعليم المفتوح) مثل Coursera و (edXوفهمها بالكامل بلغتهم الأم، مما يرفع من كفاءة التحصيل العلمي ويختصر وقت التعلم.

- **المؤسسات الأكاديمية والجامعات :** يساعد الجامعات والمراكز التعليمية في العالم العربي على توطين (Localization)المناهج والمساقات المرجعية العالمية بسرعة وكفاءة عالية، مما يساهم في إثراء المكتبات المرئية للمؤسسات التعليمية المحلية بمحتوى عالي الجودة.
- **منصات صناعة المحتوى التعليمي الرقمي :** يوفر لأصحاب القنوات والمواقع التعليمية أداة مؤتمتة، سريعة، ومنخفضة التكلفة لدبلجة مقاطعهم إلى اللغة العربية بجودة مرئية وصوتية احترافية، مما يسمح لهم بتوسيع قاعدة جمهورهم المستهدف.
- **رواد التعلم الذاتي :** يخدم الأفراد الراغبين في تطوير مهاراتهم التقنية والمهنية عبر الإنترنت، والذين يشكل ضعف الإلمام باللغة الإنجليزية عائقاً أساسياً أمام استيعابهم للمفاهيم العلمية المعقدة.

1.5 النتائج والإسهامات

قدم البحث حلولاً تقنية مبتكرة في مجال المعالجة الصوتية والمرئية، حيث تم بناء وتدريب نموذج U-Net عصبوني عميق مخصص لفصل الضجيج، مما ساهم بفاعلية في تحسين نقاء الإشارة الصوتية. كما تم تحقيق محاذاة زمنية فائقة الدقة وفصل المتحدثين وتحديد جنسهم عبر دمج تقنيات WhisperX وWav2Vec2، مما سمح بتخصيص الأصوات المترجمة بدقة .وفي الجانب البصري، تم تطبيق معمارية MuseTalk لمزامنة الشفاه، مع المساهمة بابتكار خوارزمية تنعيم رياضية (Smoothing Logic) تعتمد على فلتر الوسيط لمعالجة اهتزاز إحداثيات الوجه (Jittering) ، وتوظيف نموذج CodeFormer لترميم ملامح الوجه بمستوى موثوقية w=0.95 أظهر النظام كفاءة عالية، حيث أثبتت التجارب القدرة على دبلجة فيديو مدته 8 دقائق خلال 12 دقيقة فقط باستخدام معالجاتGPU

1.6 هيكلية البحث

يتألف هذا البحث من ستة فصول مترابطة؛ يتناول الفصل الأول المقدمة وأهمية البحث، بينما يستعرض الفصل الثاني الأدبيات السابقة حول أنظمة معالجة الصوت ومزامنة الشفاه. ويركز الفصل الثالث على المنهجية المستخدمة بدءاً من تصميم نموذج U-Net وصولاً إلى خوارزميات WhisperX ومزامنة MuseTalk وترميم CodeFormer وفي الفصل الرابع، يتم عرض نتائج التجارب ومقاييس الأداء وتحليل كفاءة النظام بالتفصيل. أما الفصل الخامس، فيحتوي على وصف التنفيذ وواجهة المستخدم، ليختتم البحث بالفصل السادس الذي يحوي الخاتمة والتوصيات المستقبلية

الفصل الثاني
الأسس التقنية للدبلجة
والأدبيات السابقة

الفصل الثاني –الأسس التقنية للدبلجة والأدبيات السابقة

2.1 المفاهيم الأساسية

قبل الخوض في التفاصيل التقنية، لا بد من شرح المفاهيم الأساسية التي يقوم عليها هذا البحث:

- **الإشارة الصوتية (Audio Signal) :** تمثيل فيزيائي للأصوات في النطاق الزمني أو الترددي، وتعتمد جودة معالجتها في أنظمة الذكاء الاصطناعي على معدل العينة (Sample Rate) ونقاء الموجة من الضجيج.
- **الترجمة (Translation) :** عملية نقل المعنى من اللغة المصدر إلى اللغة الهدف، وفي سياق هذا البحث، نركز على الترجمة السياقية (Contextual Translation) التي تحافظ على المصطلحات العلمية.
- **الدبلجة (Dubbing) :** هي الفن التقني لاستبدال المسار الصوتي الأصلي بمسار صوتي آخر بلغة مختلفة، مع الحفاظ على التزامن الزمني، هوية المتحدث، وحس الواقعية البصرية.
- **فصل المتحدثين (Speaker Diarization) :** هي عملية تقسيم الإشارة الصوتية إلى مقاطع بناءً على هوية المتحدث، والإجابة على سؤال "من يتحدث ومتى؟"، وهي خطوة أساسية في دبلجة الحوارات المتعددة .
- **مزامنة الشفاه (Lip-syncing) :** عملية توليد حركة شفاه اصطناعية تتطابق بدقة مع المسار الصوتي المولد، لضمان التوافق البصري بين الشخصية والصوت الجديد .

2.2 تطور الدبلجة من التقليدية إلى الذكاء الاصطناعي

بدأت الدبلجة في ثلاثينيات القرن الماضي مع بدايات السينما الناطقة كعملية يدوية معقدة تتطلب استوديوهات ضخمة وممثلين صوتيين ومعدات هندسية باهظة الثمن.

- **الدبلجة التقليدية :** تعتمد على العنصر البشري بشكل كامل في الترجمة، الأداء الصوتي، والمزامنة، مما يجعلها مكلفة وبطيئة.
- **الدبلجة الرقمية :** ظهرت مع تطور برمجيات تحرير الصوت، مما سهل عملية المزامنة لكنها ظلت تعتمد على الأداء البشري.
- **الدبلجة المعتمدة على الذكاء الاصطناعي :** وهي الثورة الحالية التي مكنت من أتمتة العملية بالكامل عبر تكامل نماذج التعرف على الكلام (ASR) ، والترجمة الآلية (MT[1]) ، وتوليد الكلام (TTS[2]) ، والمزامنة البصرية (Lip-sync)، مما وفر حلولاً سريعة ومنخفضة التكلفة.

[1] **MT (Machine Translation)** : الترجمة الآلية؛ وهي تقنية تعتمد على الحاسوب لترجمة النصوص أو الكلام من لغة إلى أخرى سياقياً دون تدخل بشري.

[2] **TTS (Text-to-Speech)** : توليد الكلام من النص؛ وهي تقنية برمجية تحول النصوص المكتوبة إلى مسار صوتي منطوق يحاكي النبرة البشرية طبيعياً

2.3 علاقة الذكاء الاصطناعي بمجال الدبلجة

يمثل الذكاء الاصطناعي العمود الفقري لأنظمة الدبلجة الحديثة؛ حيث يلعب دوراً حاسماً في:

1 .**عزل وتحسين الإشارة :** عبر استخدام شبكات عصبونية مثل U-Net لتنقية الصوت من الضجيج

2. **تحليل المتحدثين :** استخدام تقنيات تحديد الهوية والجنس لضمان اختيار أصوات (TTS) ملائمة لكل شخصية في الفيديو .

3. **المحاذاة الزمنية والمزامنة البصرية :** استخدام تقنيات المحاذاة القسرية (Forced Alignment) لضبط التوقيت، ونماذج الرؤية الحاسوبية لمزامنة حركة الشفاه وترميم ملامح الوجه لرفع الواقعية.

2.4 الأدبيات السابقة

تم في هذا البحث مراجعة وتحليل ما يقارب **25 دراسة مرجعية وثيقة الصلة** ، غطت الفترة الزمنية الممتدة بين عامي **2020 و2024**، والتي واكبت الطفرة النوعية المتسارعة في نماذج التعلم العميق وتوليد الوسائط المتعددة. وتنوعت هذه الجهود البحثية في بناء أنظمة الدبلجة الآلية ومعالجة الكلام عبر تكامل تقنيات التعرف على الكلام، عزل الصوت، والمزامنة البصرية. وفيما يلي استعراض لأهم الدراسات التي شكلت الأساس العلمي والتقني لهذا المشروع:

- **دراسة "WhisperX: Time-accurate speech transcription of long-form audio"** تناولت تطوير نظام WhisperX كتحسين جوهري لنموذج Whisper الأصلي؛ حيث اعتمدت على استراتيجية "القطع والدمج (Cut & Merge) المعتمدة على كاشف النشاط الصوتي (VAD) ونموذج المحاذاة القسرية لل phoneme (Forced Phoneme Alignment) لضمان توقيتات دقيقة على مستوى الكلمة، مما عالج مشكلات "الهلوسة" والانزياح الزمني محققة تسارعاً يصل إلى 12 ضعفاً [1]
- **دراسة "Pyannote.audio: neural building blocks for speaker diarization"** استعرضت نظام pyannote كمنهجية متقدمة لتقسيم المتحدثين. وفي مشروعنا، تم توظيف هذا النظام ليس فقط للفصل بين الأصوات، بل لدمجه مع نماذج Wav2Vec2 لتحديد جنس المتحدث آلياً وتخصيص البصمة الصوتية المناسبة [2] .
- **دراسة "Language-guided audio-visual source separation via trimodal consistency"** بحثت في تقنيات عزل المصادر مثل نظام VAST الذي يعتمد على الإشراف الذاتي، وفي هذا السياق، قمنا ببناء نموذج عزل الضوضاء (Speech-Noise Separation) الخاص بنا لتعزيز جودة الإشارة وتنقيتها قبل مرحلة الدبلجة، مما يضمن نقاء صوت المتحدث الوحيد ووضوحه [3]
- **دراسة "XTTS: a massively multilingual zero-shot text-to-speech model"** ركزت على نموذج XTTS لاستنساخ البصمة الصوتية (Voice Cloning) بدقة عالية، ومع ذلك، اعتمد مشروعنا على تقنية Edge TTSلتفوقها في جودة اللغة العربية وسرعة الاستجابة (Low Latency) وتوفير نبرات طبيعية دون الحاجة لمتطلبات عتادية ضخمة [4]
- دراسة **"MuseTalk: Real-Time High Quality Lip Synchronization with Latent Space Inpainting"** قدمت نموذج **MuseTalk**الذي يعتمد على آلية "ترميم الفضاء الكامن (Latent Space Inpainting) " عبر شبكة VAE للتنبؤ بحركة الفم، واستخدام معمارية U-Net متعددة المقاييس لدمج الميزات الصوتية والبصرية، ويمثل هذا النموذج المعيار التقني الأساسي المعتمد في مشروعنا لضمان توليد حركة شفاه واقعية للغاية ومتطابقة تماماً مع ال phoneme المستخلصة وبسرعة فائقة توافق الوقت الحقيقي (Real-Time) دون تغيير الهوية الأصلية للوجه **[5]**

ويمكن تلخيص الدراسات المرجعية في الجدول التالي :

جدول (1-2) :

مقارنة تقنية بين منهجيات ومعايير قياس الأبحاث السابقة

المرجع	التقنية (Technology)	آلية العمل (Methodology)	معايير التقييم (Metrics)	الداتا المستخدمة (Datasets)	العام
[1]	WhisperX	يعتمد استراتيجية Forced Phoneme Alignment عبر دمج نموذج Whisper مع نماذج التمثيل الصوتي لربط (wav2vec2 مثل) النص بالصوت بدقة زمنية عالية.	WER (نسبة خطأ الكلمات)، RTF (عامل الوقت الحقيقي، ودقة الطوابع الزمنية	LibriSpeech, VoxPopuli, ومجموعة بيانات Whisper الضخمة.	2023
[2]	pyannote.audio	بنية عصبونية تعتمد على EEND-VC لتقسيم المتحدثين، وتستخدم تقنيات العنقدة (Clustering) لتحديد هوية كل متحدث في المقطع.	DER (معدل خطأ التقسيم)، JER (معدل خطأ jaccard error rate لتقييم دقة تحديد هوية المتحدث	AMI Corpus, VoxConverse, DIHARD	2020
[3]	VAST	يعتمد على Trimodal Consistency (الاتساق الثلاثي) بين الصوت، اللغة، والرؤية عبر تقنية التعلم التبايني (Contrastive Learning) لعزل المصادر.	SDR (نسبة الإشارة إلى التشوّه)، CLAPScore لتقييم الصلة الدلالية بين النص والصوت المعزول.	VGGSound, AudioCaps, Clotho	2023
[4]	XTTS	نموذج لغوي (GPT-based) يعتمد على ترميز الصوت (Audio Tokenization) واستنساخ النبرة (Voice Cloning) لتوليد كلام طبيعي بلغات متعددة.	CER (نسبة خطأ الحروف) SECS (تشابه المتحدث)، MOS (مقياس جودة الصوت الطبيعي)	Multilingual LibriSpeech (MLS), Common Voice, LibriTTS	2024

2024	HDTF, MEAD, VFHQ	FID (جودة الصور وتفاصيل الوجه) CSIM (ثبات ملامح الهوية) LSE-C (ثقة مزامنة الشفاه)	يعتمد على ترميم الفضاء الكامن (Latent Space Inpainting) بالتنبؤ بحركة الفم داخل فضاء الكامن، ودمج ميزات VAE الصوت والصورة عبر شبكة U-Net متعددة المقاييس لإنتاج مزامنة فورية عالية الجودة.	MuseTalk	**[5]**

الفصل الثالث

التقنيات والأدوات المستخدمة

الفصل الثالث – التقنيات والأدوات المستخدمة

3.1 المفاهيم التقنية والخصائص الأساسية :

في هذا الجزء، نناقش التقنيات الجوهرية التي شكلت العمود الفقري للنظام، وكيف تم توظيف كل منها بشكل تخصصي لرفع كفاءة الدبلجة:

3.1.1 تقنية Demucs لفصل المصادر الصوتية :

- **ما هوDemucs ؟** هو نموذج ذكاء اصطناعي متطور طورته أبحاث شركة(Meta AI) ، يعتمد على معمارية-U" "Net و "Transformers" صُمم خصيصاً لفصل الإشارة الصوتية إلى مسارات منفصلة بدقة عالية.

- **دوره في المشروع :** يعمل **Demucs** كخبير متخصص في **فصل الموسيقى والآلات الخلفية** .في الفيديوهات التعليمية التي تحتوي على موسيقى تصويرية أو مقدمات موسيقية، يتم استخدام هذا النموذج لعزل صوت الموسيقى تماماً عن صوت الكلام، مما يضمن الحصول على مسار بشري نقي قبل البدء بالترجمة.

3.1.2 نموذج ResUNet المطور (النموذج الخاص بنا) :

- **وظيفته :** هذا هو النموذج الذي تم بناؤه وتطويره ضمن هذا المشروع، وهو متخصص في **فصل الضجيج البشري والبيئي.**

- **دوره في المشروع :** على عكسDemucs ، يركز نموذجنا على عزل "الأصوات البشرية المتداخلة(Chatter) "، أو ضجيج القاعات، أو التشويش الناتج عن الميكروفون. هذا التكامل بين النموذج الخاص بنا و Demucs يسمح للنظام بالتعامل مع أي نوع من التشويش، سواء كان "موسيقياً" أو "بشرياً/بيئياً."

3.1.3 تقنية WhisperX للمحاذاة والتعرف على الكلام :

- **ما هوWhisperX ؟** هو نسخة "محسنة" ومطورة من نموذج Whisper الأصلي، يمثل نظاماً متكاملاً للمحاذاة الزمنية الدقيقة.(Time Alignment)

- **دوره في المشروع :** يتولى WhisperX مهمة تحويل الكلام المسموع إلى نص مترجم مع تحديد "طوابع زمنية " (Timestamps)دقيقة على مستوى الكلمة وحتى مستوى "ال phoneme" (أصغر وحدة صوتية) .

- **الإضافة النوعية (فصل المتحدثين) :** يدمج النظام نموذج **Pyannote** ضمن WhisperX للقيام بعملية فصل المتحدثين (Speaker Diarization)، مما يسمح بتحديد هوية كل متحدث وزمن حديثه بدقة في الفيديوهات التي تحتوي على أكثر من شخص.

3.1.4 المقارنة بين Whisper (الأصلي) و WhisperX (المطور) :

توضح المقارنة التالية سبب تفضيلنا لـ **WhisperX** لضمان جودة الدبلجة الاحترافية:

جدول (1-3) :

مقارنة بين whisper & whisperx

وجه المقارنة	Whisper (OpenAI)	WhisperX (المطور)
دقة التوقيت	يعطي توقيت تقريبي للجملة (قد يتأخر الصوت عن الصورة).	يعطي توقيت دقيق جداً **على مستوى الكلمة** (Word-level).
السرعة	أبطأ في المعالجة التسلسلية.	**أسرع بـ 10 مرات** بفضل تقنية المعالجة بالدفعات (Batched Inference).
المحاذاة (Alignment)	لا يمتلك خاصية المحاذاة الصوتية.	يستخدم نماذج (مثل Wav2Vec2) لعمل **محاذاة لفظية دقيقة**
تعدد المتحدثين	يجد صعوبة في التمييز بين المتحدثين.	يدعم خاصية (Diarization) لمعرفة "من قال ماذا".
الصمت والضجيج	قد يهلوس (Hallucination) عند وجود صمت طويل.	يستخدم VAD (كاشف نشاط الصوت) لتجاهل الصمت بدقة.

3.1.5 تقنية Pyannote.audio لفصل المتحدثين :

ما هوPyannote ؟ هو مكتبة ذكاء اصطناعي مفتوحة المصدر متخصصة في تحليل الإشارات الصوتية لحل مسألة "من يتحدث ومتى؟" بدقة عالية

دوره في المشروع :يعمل كخبير تحليل الهوية الصوتية؛ حيث يتدخل ليقوم بتقسيم الفيديو إلى مقاطع زمنية مفصولة بناءً على بصمة صوت كل متحدث، مما يسمح للنظام بالتعامل مع الفيديوهات التعليمية التي تضم أكثر من شخص بشكل احترافي.

3.1.6 تقنية Wav2Vec2 لتصنيف جنس المتحدث :

وظيفتها : تُستخدم هذه التقنية المعتمدة على نماذج التصنيف الصوتي لتحديد جنس المتحدث (ذكر/أنثى) آلياً

دورها في المشروع :تعد هذه الخطوة محورية في أتمتة الدبلجة، حيث يقوم النظام بناءً على التصنيف باختيار صوت الـ (TTS) المناسب لجنس كل متحدث في الفيديو، مما يضمن توافق الهوية الصوتية المترجمة مع الشخصية الظاهرة.

3.1.7 تقنيات الترجمة اللغوية (LLM) والمعالجة النصية :

ما هي؟ استخدام نماذج لغوية ضخمة (LLMs) للترجمة السياقية، مقترنة بخوارزميات لمعالجة النصوص (Text Normalization)

دورها في المشروع : تتولى ترجمة النص المستخلص من الإنجليزية إلى العربية مع الحفاظ على الدقة العلمية، كما تقوم بتحويل الأرقام والرموز إلى كلمات عربية مكتوبة لضمان نطقها آلياً بشكل صحيح وسليم.

3.1.8 تقنيات توليد النطق والاستنساخ :

ما هي؟ نماذج متطورة لتوليد الصوت البشري) مثل Edge TTS أو (ElevenLabs وتقنيات الاستنساخ الصوتي) مثل XTTS v2)

دورها في المشروع : توليد الصوت العربي بناءً على الجنس المحدد، مع تطبيق خوارزمية (Atempo) لضبط سرعة النطق ليتطابق زمنياً مع الفيديو الأصلي، وتوحيد مستويات الصوت والدمج المتقدم (Mixing & Ducking)

3.1.9 معمارية MuseTalk للمزامنة البصرية :

ما هي؟ هي معمارية توليدية متطورة تعمل في الفضاء الكامن (Latent Space) لمزامنة حركة الشفاه مع أي مقطع صوتي مدخل

دورها في المشروع : تتولى مهمة تعديل حركة فم المتحدث في الفيديو الأصلي ليتطابق بصرياً مع الكلام العربي المولد، مما ينقل النظام من مرحلة "تركيب الصوت" إلى مرحلة "الدبلجة البصرية" المتكاملة.

3.1.10 معمارية CodeFormer لترميم وتحسين معالم الوجه :

وظيفتها : نموذج تعزيز بصري يعتمد على المحولات (Transformers) لترميم ملامح الوجه وإزالة التشوهات الناتجة عن عمليات المعالجة الرقمية

دورها في المشروع : تُستخدم كخطوة معالجة نهائية لضمان جودة سينمائية؛ حيث تم ضبط معامل الموثوقية (Fidelity) عند مستوى w=0.95 للحفاظ على هوية الشخصية الأصلية مع رفع دقة الملامح والشفاه المولّد

الفصل الرابع
منهجية البحث والمنظومة المقترحة

الفصل الرابع– منهجية البحث والمنظومة المقترحة

يستعرض هذا الفصل المنهجية العلمية المتبعة في بناء وتصميم نظام الدبلجة الآلي المقترح. يبدأ الفصل بتوصيف مجموعة البيانات المستخدمة وكيفية تحضيرها، يليه المخطط الانسيابي العام للمنظومة، ومن ثم الانتقال إلى الشرح التفصيلي والمطوّل لكافة مراحل الPipeline من المعالجة الصوتية وحتى المعالجة البصرية، ليختتم الفصل بالتحليل الهندسي للشبكة العصبونية (ResUNet) المخصصة لعزل الضجيج

4.1 مجموعة المعطيات المستخدمة وتحضيرها

لضمان كفاءة وثبات النموذج في ظروف التشغيل الواقعية، تم الاعتماد على مجموعة بيانات **VoiceBank-DEMAND** المعيارية .(Benchmark Dataset) تتميز هذه المجموعة بتغطية واسعة ودقيقة لأنواع الضجيج المختلفة المسجلة في بيئات حقيقية (مثل ضجيج المكاتب، وسائل النقل، والشارع)، حيث تم دمجها برمجياً مع كلام بشري نقي مستخلص من قاعدة بيانات **VCTK**العالمية.

يوضح الجدول (4-1) التوصيف الإحصائي الموسع وتوزيع البيانات المستخدمة في مرحلتي التدريب والاختبار:

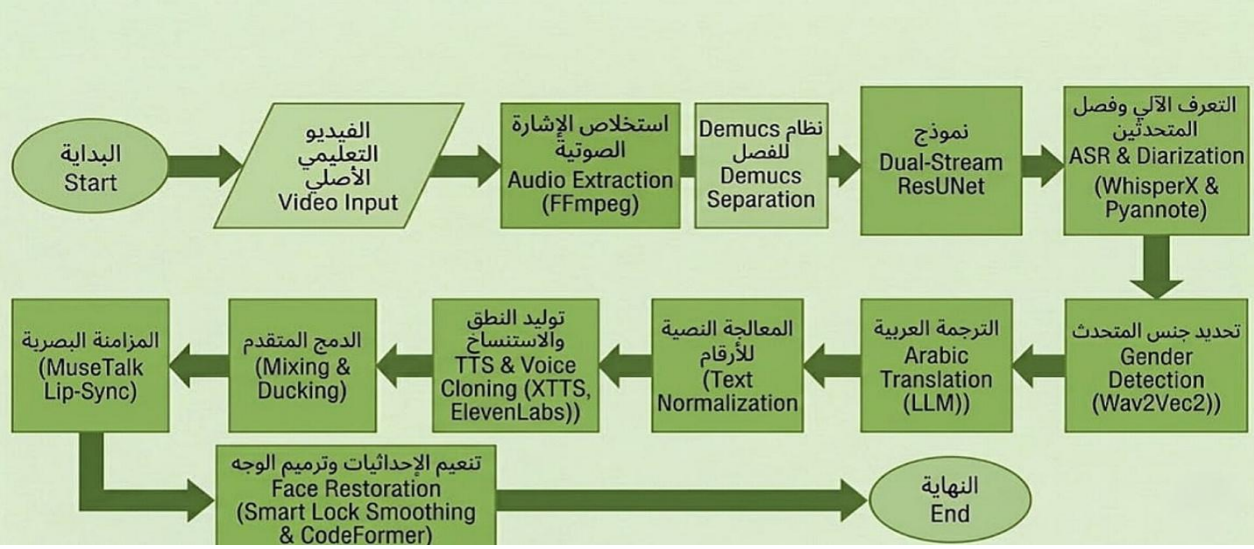
جدول (4-1):

إحصائيات تقسيم مجموعة البيانات المستخدمة

المعيار	التفاصيل والكمية
إجمالي عدد الملفات الصوتية	أكثر من 12,000 مقطع صوتي موزعة بين التدريب والاختبار.
إجمالي المادة الصوتية	ما يقارب 10 ساعات من التسجيلات الصوتية المتنوعة.
عدد المتحدثين (Training)	28 متحدثاً (14 ذكور، 14 إناث) لضمان تنوع النبرات.
عدد المتحدثين (Testing)	متحدثان معزولان تماماً لضمان حيادية التقييم.
معدل العينة (Sampling Rate)	موحد عند 16,000 هرتز لضمان استقرار المعالجة التلافيفية.

4.2 مخطط المنهجية العام :

يعتمد نظام الدبلجة الآلي المقترح على Pipeline متسلسل ومتكامل، يربط بين معالجة الإشارات الصوتية الرقمية ونماذج التوليد البصري. يوضح **الشكل (1-4)** المخطط الانسيابي العام لكيفية تدفق البيانات داخل المنظومة، بدءاً من إدخال الفيديو الخام باللغة الإنجليزية ووصولاً إلى تصدير الفيديو المدبلج باللغة العربية بجودة سينمائية احترافية.



الشكل (1-4): المخطط الانسيابي العام لمنهجية عمل نظام الدبلجة الآلي

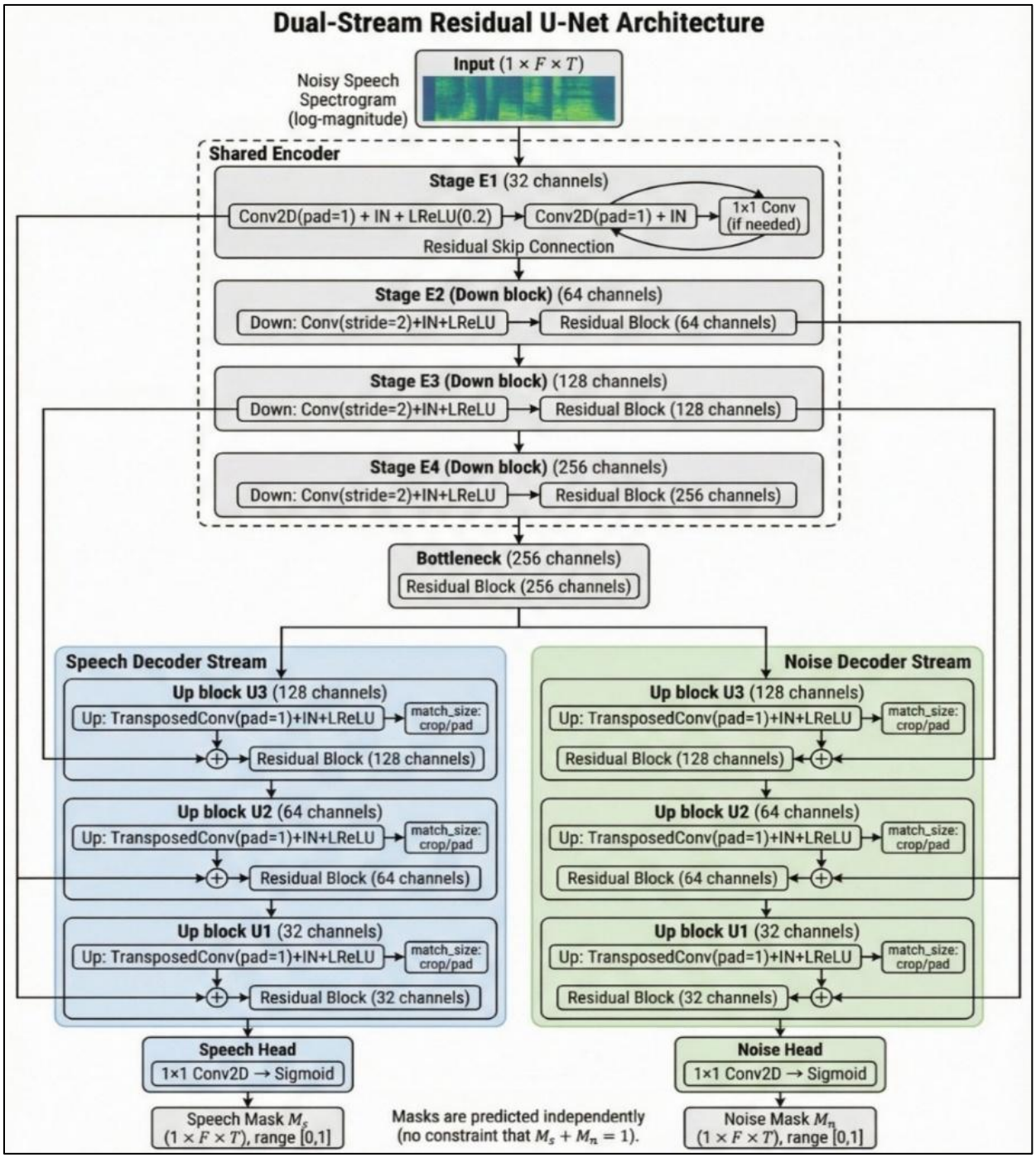
وتتلخص وظائف المكونات المنهجية فيما يلي:

- **استخلاص الإشارة الصوتية (Audio Extraction) :** باستخدام أداةFFmpeg ، يتم فصل المسار الصوتي عن الفيديو الأصلي بصيغة WAV وبمعدل عينة 16 kHz لضمان توافق البيانات مع النماذج اللاحقة.
- **نظام Demucs للفصل (Demucs Separation) :** تعمل هذه المرحلة على عزل الموسيقى الخلفية والآلات الموسيقية عن المسار الصوتي، مما يترك كلاماً بشرياً أولياً يحتاج لمزيد من التنقية.
- **نموذج (Dual-Stream ResUNet) :** وهو المكون الجوهري المطور في البحث، حيث يقوم بتنقية الكلام من الضجيج البيئي والبشري والصدى، لإنتاج إشارة صوتية نقية تماماً جاهزة لمرحلة التعرف.
- **التعرف الآلي وفصل المتحدثين (WhisperX & Pyannote) :** يتم تحويل الصوت النقي إلى نص، بالتوازي مع استخدام نموذج Pyannote المدمج لفصل المتحدثين (Speaker Diarization) وتحديد الطوابع الزمنية الدقيقة لكل متحدث على مستوى الكلمة.
- **تحديد جنس المتحدث (Gender Detection) :** يمر الصوت البشري عبر نموذج تحليل النبرة المعتمد على Wav2Vec2 لتحديد الجنس (ذكر/أنثى) بناءً على نظام التصويت لأطول المقاطع الصوتية، مع توفير تحليل التردد الأساسي (Pitch fallback) كبديل احتياطي.
- **الترجمة العربية (Arabic Translation) :** استخدام نماذج لغوية ضخمة (LLM) لترجمة النص المستخلص من الإنجليزية إلى العربية مع الحفاظ على الدقة العلمية والسياق التعليمي لضمان جودة المحتوى.
- **المعالجة النصية للأرقام (Text Normalization) :** تحويل الأرقام والرموز الإنجليزية ضمن النص المترجم إلى كلمات عربية مكتوبة لضمان نطقها آلياً بشكل صحيح وسليم.

- **توليد النطق والاستنساخ (TTS & Voice Cloning) :** بناءً على الجنس، يتم استخدام نماذج التوليد مثل (ElevenLabs) مع الاعتماد على الاستنساخ الصوتي (XTTS v2) للحصول على نبرة واقعية، وتطبيق خوارزمية لضبط سرعة النطق (Atempo) لتتوافق مع المدة الزمنية الأصلية.
- **الدمج المتقدم (Mixing & Ducking) :** دمج النطق العربي مع الموسيقى الخلفية، مع تطبيق خفض تلقائي للموسيقى عند بدء الكلام (Sidechain Ducking) وتوحيد مستويات الصوت.
- **المزامنة البصرية (MuseTalk Lip-Sync) :** يتم إدخال إطارات الفيديو مع ملف الصوت العربي المولد إلى معمارية MuseTalk لتوليد حركة شفاه متطابقة تماماً مع ال phoneme العربية المستخلصة في الفضاء الكامن.
- **تنعيم الإحداثيات وترميم معالم الوجه (Smoothing & CodeFormer) :** يمر الفيديو المولد عبر خوارزمية التنعيم الذكية الخاصة بالنظام (Smart Lock Smoothing) القائمة على تطبيق "فلتر الوسيط (Median Filter)" بحجم نافذة زمنية يساوي 15 إطاراً لإلغاء الرجفة، يليه نموذج CodeFormer لترميم ملامح الوجه بدقة فائقة وبمعامل موثوقية w=0.95 لاستعادة التفاصيل السينمائية دون تغيير الهوية الأصلية.

4.3 التحليل الهندسي لبنية نموذج ResUnet المطور

يعتمد النموذج على بنية مشفر وفك تشفير (Encoder-Decoder) من نوع ResUNet ، مصممة لتحقيق توازن مثالي بين استعادة جودة الصوت وسرعة المعالجة .ويوضح **الشكل (2-4)** الهيكلية التفصيلية لهذه البنية، حيث نلاحظ كيف يقوم المشفر باستخلاص الميزات بينما تعمل طبقات فك التشفير على إعادة بناء الإشارة الصوتية النقية مع الحفاظ على التفاصيل عبر وصلات التخطي.



الشكل (2-4): بنية ResUnet

وتتكون هذه البنية من:

- **الطبقات التلافيفية المتبقية (Residual Convolutional Layers) :**
  - تستخدم لاستخراج الميزات العميقة من بيانات المخطط الطيفي عبر مرشحات متصاعدة (32, 64, 128)
  - تعتمد على دالة التنشيط **Leaky ReLU** لضمان تدفق التدرجات ومنع مشكلة "تلاشي التدرج" في الطبقات العميقة.
- **وصلات التخطي (Skip Connections) :**
  - تعمل على نقل المعلومات المكانية والترددية من المشفر مباشرة إلى فك التشفير، مما يحافظ على التفاصيل الدقيقة للصوت البشري ويمنع فقدان المعلومات أثناء الضغط.
- **طبقات التقييس والانتظام (Batch Normalization & Dropout) :**
  - تستخدم **Batch Normalization** لتحقيق استقرار التدريب وسرعة التقارب.
  - تستخدم طبقات **Dropout** بنسبة محددة لمنع التكيف الزائد(Overfitting) ، مما يجعل النموذج قادراً على التعامل مع أنواع ضجيج لم يراها مسبقاً.
- **طبقة إعادة البناء (Decoder & Upsampling) :**
  - تقوم بإعادة بناء الإشارة الصوتية النقية، وتستخدم في نهايتها طبقة تلافيفية لتوليد **القناع (Mask)** الذي يُضرب في الإشارة الأصلية لاستخلاص الصوت النقي

4.4 التحسينات التي قدمها هذا المنهج مقارنة بالأساليب التقليدية :

1. **تحسين دقة الفصل (Enhanced Accuracy) :** القدرة على استخلاص ميزات مكانية وترددية معقدة، مما يزيد من دقة التعرف على الكلام في الفيديوهات المشوشة.
2. **تقليل أخطاء التزامن (Reduced Misclassifications) :** دمج **WhisperX** يضمن مطابقة النص المترجم مع الطوابع الزمنية للفيديو بدقة "ال phoneme"، مما يلغي التداخل الزمني.
3. **سرعة الأداء (Optimized Speed) :** المعالجة المباشرة للمخططات الطيفية تتيح دبلجة فيديو مدته 8 دقائق في غضون 12دقيقة عند استخدام.GPU
4. **بنية أكثر استقراراً :** استخدام طبقات **Dropout** و**Batch Normalization** يضمن بقاء النموذج فعالاً حتى مع فيديوهات تعليمية ذات جودة تسجيل ضعيفة.

الفصل الخامس

التجارب والنتائج والتقييم

الفصل الخامس– التجارب والنتائج والتقييم

يستعرض هذا الفصل الجوانب التطبيقية والمنهجية لنظام الدبلجة الآلي المقترح، مع التركيز على آليات التنفيذ وقياس كفاءة الأداء. يتناول الفصل التحديات التقنية المواجهة، وتوصيفاً دقيقاً لمنهجية العمل والبيانات المستخدمة، ليختتم بعرض مقاييس التقييم والنتائج الإحصائية التي تثبت فاعلية النظام في بيئة الاختبار.

5.1 التحديات والاحتياجات :

قبل البدء بتنفيذ النظام، واجهنا مجموعة من التحديات الجوهرية التي تطلبت استراتيجيات معالجة خاصة:

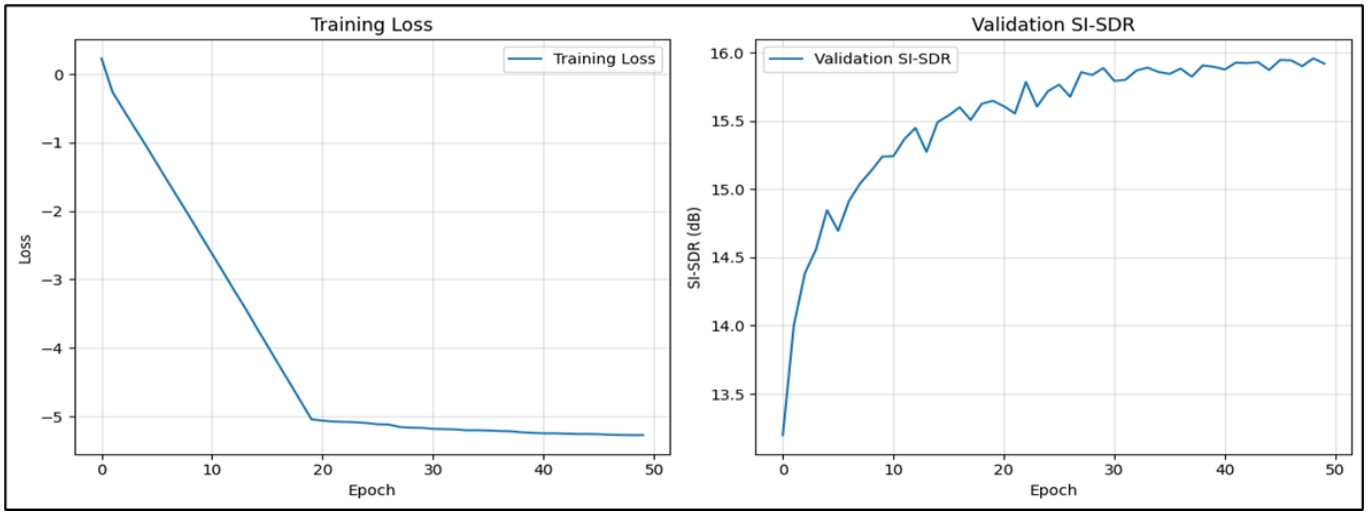
- **مجموعات المعطيات :** تطلبت الشبكات التلافيفية بيانات ضخمة ومتنوعة للتكيف مع أنماط الضجيج المختلفة.
- **تعقيد الإشارة وتعدد المتحدثين :** تداخل الأصوات التعليمية مع الموسيقى أو التشويش البيئي، بالإضافة إلى وجود أكثر من متحدث في بعض المقاطع، مما تطلب آلية لفصل الهويات الصوتية وتحديد جنس المتحدث لتوجيه نظام توليد الصوت بدقة
- **التفاعل مع البيئة :** الحاجة لتطوير قدرات النموذج ليتكيف مع اختلاف جودة الميكروفونات في التسجيلات التعليمية.
- **مشكلة عدم الاستقرار البصري (Jittering) :** أثناء توليد حركة الشفاه، ظهرت مشكلة اهتزاز ورجفة في إحداثيات الوجه (Bounding Boxes) وتم التغلب عليها برمجياً بافتراض حل رياضي متقدم يعتمد على فلاتر التنعيم لحفظ الاستقرار الزمني للإطارات.

5.2 آلية عمل الاختبار والتدريب :

تم تنفيذ التجارب في بيئة Kaggle باستخدام معالج Tesla T4 GPU ، مع اتباع الاستراتيجيات التالية:

- **المحسن :** استخدام AdamW لضبط الأوزان مع تنظيم فعال للوزن.
- **جدولة التعلم :** تفعيل تقنية Learning Rate Warmup لضمان استقرار الأوزان في البداية.
- **المعالجة المسبقة :** توحيد معدل العينة إلى 16 kHz وتحويل الموجات إلى نطاق التردد (STFT)

ويوضح **الشكل (5-1)** منحنيات الأداء أثناء عملية التدريب؛ حيث نلاحظ الانخفاض التدريجي والمستقر في دالة الخسارة (Training Loss)، بالتوازي مع الارتفاع الملحوظ في مقياس (SI-SDR) لمجموعة التحقق، مما يثبت وصول النموذج إلى حالة التقارب (Convergence) المنشودة دون حدوث إفراط في التدريب (Overfitting) خلال الـ 50 حقبة المحددة.



الشكل (1-5): منحنيات تطور دالة الخسارة ومقياس SI-SDR أثناء تدريب النموذج

5.3 مقاييس التقييم :

للحكم على جودة النظام وفصل الصوت، تم اعتماد مجموعة من المقاييس الرياضية المعيارية. يوضح **الجدول (2-4)** المقاييس المستخدمة والهدف من كل منها، حيث تعطي هذه المقاييس مؤشراً دقيقاً عن مدى وضوح الكلام ونسبة الخطأ اللغوي

جدول (1-5) : مقاييس تقييم الأداء المستخدمة في المشروع

المقياس	الوصف والهدف
PESQ	لتقييم جودة الصوت الإدراكية ومدى قربه من الصوت النقي الأصلي.
STOI	لقياس مدى وضوح وفهم الكلام البشري بعد المعالجة.
WER	لقياس دقة التعرف على الكلام عبر حساب نسبة الكلمات الخاطئة.
DNSMOS	مقياس موضوعي لجودة فصل الضجيج في البيئات الواقعية.

5.4 نتائج الاختبار وتحليل الأداء :

بعد إتمام مراحل التدريب والمعالجة، تم تقييم أداء النظام باستخدام عينة الاختبار من مجموعة بيانات-VoiceBank DEMAND يوضح **الشكل (2-5)** القيم الإحصائية التي حققها النظام عبر مقاييس الجودة العالمية ؛ حيث تعكس هذه الأرقام كفاءة النموذج في الحفاظ على جودة الإشارة الصوتية وتقليل نسبة الخطأ في الترجمة والمحاذاة.

```
==================================================
🏆 RESULTS:
==================================================
🔹 Mean PESQ: 2.6918
🔹 Mean STOI: 0.9385
🔹 Mean DNSMOS: 3.0857
🔹 Mean WER: 0.0219
🔹 Mean WER_NOISY: 0.0241
🔹 Mean SI_SDR: 20.2707
```

الشكل (2-5): النتائج الإحصائية النهائية لأداء نظام الدبلجة الآلي

ويمكن تحليل دلالة هذه القيم وفقاً لما يلي:

مقياس الجودة الإدراكية الصوتية (Mean PESQ) : بلغت القيمة المحققة 2.6918، وهي تشير إلى جودة إدراكية إجمالية تقع في النطاق (Fair to Good) وفقاً للمعيار العالميITU-T P.862 ، مما يعكس كفاءة النموذج في إزالة المؤثرات الخارجية مع الحفاظ على البنية الأساسية للصوت البشري.

مقياس وضوح ومفهومية الكلام (Mean STOI) : تم تسجيل قيمة مرتفعة بلغت 0.9385، وتُعد هذه النتيجة (المقتربة من 94%) مؤشراً ممتازاً يثبت أن النظام يحافظ تماماً على مفهومية الكلام ومخارج الحروف دون حدوث أي بتر أو تشويه يمنع استيعاب الكلمات المنطوقة.

المقياس الذكي لإلغاء الضجيج (Mean DNSMOS) : بلغت القيمة 3.0857، وهي قيمة تتجاوز المتوسط وتؤكد نجاح الشبكة العصبية العميقة في خفض مستويات الضجيج في الخلفية بأسلوب يحاكي تقييم المستمع البشري.

معدل الخطأ في التعرف على الكلمات (WER) : أظهرت مقارنة المعايير تحسناً في كفاءة معالجة النصوص؛ حيث انخفض معدل الخطأ في التعرف الآلي على الكلام من 0.0241 للصوت الأصلي المشوش (Mean WER_NOISY) إلى 0.0219 بعد تطبيق نموذج التحسين الحالي(Mean WER) ، مما يثبت القيمة التطبيقية للنظام في تهيئة الصوت لخدمة أنظمة الذكاء الاصطناعي الأخرى بدقة أعلى.

نسبة الإشارة إلى التشوه الثابتة (Mean SI_SDR) : حقق النموذج قيمة هندسية متميزة بلغت 20.2707 .dB وبما أن القيم المعيارية التي تتخطى حاجز 15 ديسيبل تُصنف عالمياً كقيم ممتازة (Outstanding) في مهام فصل وتحسين الإشارة، فإن هذه القيمة توثق رياضياً نجاح النموذج في عزل الصوت المستهدف بنقاء عالٍ وبأقل حد ممكن من الأخطاء الرياضية الناتجة عن المعالجة

الفصل السادس

الخاتمة والتوصيات المستقبلية

الفصل السادس -الخاتمة والآفاق المستقبلية

6.1 الخاتمة :

تم في هذا البحث استخدام تقنيات تعلم عميق متقدمة لتطوير نموذج فعال لدبلجة الفيديوهات التعليمية من اللغة الإنجليزية إلى العربية دبلجةً صوتية وبصرية متكاملة؛ وذلك عبر دمج معمارية U-Net لمعالجة وفصل الضجيج الصوتي، وتقنيات WhisperX و Pyannote للمحاذاة الزمنية الدقيقة وفصل المتحدثين، وصولاً إلى تحقيق المزامنة البصرية لحركة الشفاه باستخدام MuseTalk وترميم ملامح الوجه فائقة الدقة عبر CodeFormer.

أظهرت التجارب أن النموذج يحقق أداءً متزناً في تحسين جودة الإشارة الصوتية، دقة التعرف على الكلام، والتوافق البصري، مع قدرة عالية على التعامل مع السيناريوهات التعليمية المختلفة التي تشوبها ضوضاء الخلفية وتعدد المتحدثين. وبالرغم من النجاح في استنساخ الصوت ومزامنة الشفاه، لا يزال هناك مجال للتحسين في جوانب مثل التعامل مع التداخلات الصوتية المعقدة جداً، ونقل "المشاعر والانفعالات" بدقة إلى النبرة الصوتية وحركة الوجه في الفيديوهات ذات الجودة المتدنية. يمكن لهذا النموذج المتكامل أن يشكل حجر زاوية لتطبيقات عملية في دبلجة المنصات الأكاديمية العالمية، ودعم أنظمة التعليم المفتوح، وأتمتة إنتاج المحتوى المرئي المترجم باحترافية عالية.

ويمكن تلخيص النتائج المخرجة من هذا النظام على شكل التعدادات التالية:

1. بناء نظام لـ دبلجة الفيديوهات التعليمية صوتاً وصورة من اللغة الإنجليزية إلى العربية بدقة وأداء متوازن يربط البيئة السمعية والبصرية تزامناً متكاملاً.

2. تحقيق التعرف الآلي على الكلام بدقة عالية، حيث انخفضت نسبة الخطأ (WER) إلى 0.02 فقط.

3. تم فصل وإلغاء ضجيج الخلفية بدقة ممتازة تضمن الحفاظ على مفهومية ووضوح النطق البشري بنسبة تتجاوز 93%.

4. عزل وتحسين الإشارة الصوتية بدقة هندسية مرتفعة بلغت 20.27 ديسيبل (SI_SDR)، مما يضمن نقاء الصوت المستهدف بأقل تشويه ممكن ويمكن تلخيص النتائج المخرجة من هذا النظام على شكل التعدادات التالية:

5. بناء نظام لـ دبلجة الفيديوهات التعليمية صوتاً وصورة من اللغة الإنجليزية إلى العربية بدقة وأداء متوازن يربط البيئة السمعية والبصرية تزامناً متكاملاً.

6. تحقيق التعرف الآلي على الكلام بدقة عالية، حيث انخفضت نسبة الخطأ (WER) إلى 0.02 فقط.

7. فصل وإلغاء ضجيج الخلفية بدقة ممتازة تضمن الحفاظ على مفهومية ووضوح النطق البشري بنسبة تتجاوز 93%.

8. تم عزل وتحسين الإشارة الصوتية بدقة هندسية مرتفعة بلغت 20.27 ديسيبل (SI_SDR)، مما يضمن نقاء الصوت المستهدف بأقل تشويه ممكن.

6.2 الآفاق المستقبلية :

لتعزيز هذا المشروع وتوسيع نطاق تأثيره، يمكن التركيز على الآفاق المستقبلية التالية:

•استنساخ المشاعر والتعبير الصوتي (Emotional Voice Cloning) :

- تطوير تقنيات متقدمة تتيح للنظام ليس فقط محاكاة نبرة المحاضر، بل نقل حالته العاطفية (كالحماس أو التأكيد) إلى الصوت العربي المولد، مما يساهم في جعل الدبلجة تبدو أكثر تفاعلية وطبيعية.

•توسيع قاعدة المعطيات اللغوية :

- جمع مزيد من البيانات التي تشمل اللهجات العربية المختلفة والترجمات المتخصصة في مجالات (الطب، الهندسة، القانون) لزيادة قدرة النموذج على فهم المصطلحات التقنية المعقدة بدقة أعلى.

•تحسينات بصرية متقدمة وتكامل الحركة :

- توسيع عمل النماذج البصرية لتشمل التعامل مع زوايا الوجه الجانبية (Profile Faces) وتوليد إيماءات لغة الجسد المتوافقة مع سياق الكلام، لتعزيز الاندماج البصري بين حركة الشفاه وحركة الجسم الكلية.
- تطبيق تقنيات التعلم المستمر (Continual Learning) لتكييف النموذج مع البيانات الجديدة وأنماط الوجوه المختلفة دون الحاجة لإعادة التدريب الكامل.

•تحسين عمليات المعالجة الفورية (Streaming) :

- تطوير تقنيات لمعالجة التدفقات الصوتية والبصرية مباشرة، مما يسمح بالترجمة والمزامنة والدبلجة الفورية أثناء البث المباشر للمحاضرات والندوات التعليمية بدقة تفوق الأداء الحالي.

•الأنظمة متعددة الوسائط (Multimodal Optimization) :

- استغلال البيانات البصرية (قراءة لغة الشفاه وحركة الوجه) كمدخلات إضافية (Input) لتعزيز دقة التعرف على الكلام (ASR) في حالات الضجيج الشديد، مما يحسن سياق التعرف والترجمة بشكل تكاملي.

6.3 التوصيات :

1. **تحسين جودة المعطيات :** لتعزيز دقة النموذج في مختلف الظروف، يجب التركيز على تنويع بيانات التدريب بما يعكس تعقيدات العالم الحقيقي (مثل صدى القاعات، ضجيج الأجهزة، وتفاوت جودة الميكروفونات).
2. **تطوير أدوات تحليلية مرئية :** لتسهيل تفسير أداء النظام وتحديد نقاط الضعف في ترجمة أو محاذاة كلمات محددة بشكل أكثر فعالية.
3. **تعزيز التعاون مع مجالات الذكاء الاصطناعي الأخرى :** للاستفادة من التطورات في معالجة اللغة الطبيعية (**NLP**) لدمج أنظمة دبلجة أكثر شمولية وفهماً للسياق الثقافي العربي.

6.4 الملاحظات الختامية :

يشكل هذا البحث خطوة مهمة نحو استخدام الذكاء الاصطناعي في كسر الحواجز اللغوية للوصول إلى المعرفة العالمية ومع تقدم تقنيات التعلم العميق، يمكن تحسين دقة المنظومة وجعلها أكثر كفاءة في البيئات المختلفة تشير الدراسة إلى أن التحسينات المستمرة ستساعد في تطوير أنظمة ذكية أكثر موثوقية، مما يساهم في تعزيز التواصل والشمولية التعليمية للمجتمع العربي المعتمد على المحتوى الرقمي

المراجع References

[1] Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

study link

[2] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... & Gill, M. P. (2020, May). Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7124-7128). IEEE.

study link

[3]Tan, R., Ray, A., Burns, A., Plummer, B. A., Salamon, J., Nieto, O., ... & Saenko, K. (2023). Language-guided audio-visual source separation via trimodal consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10575-10584).

study link

[4] Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., ... & Weber, J. (2024). Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

study link

[5] Zhang, Y., Minhao, L.I.U., Chen, Z., Wu, B., Zhan, C., He, Y., HUANG, J. and Zhou, W., 2024. Musetalk: Real-time high quality lip synchronization with latent space inpainting.

study link

Syrian Private University

Faculty of Engineering

Artificial intelligence and Data science

Automated Dubbing System for Videos using Deep Learning Techniques (English to Arabic)

This thesis is submitted in partial

fulfillment of the requirements for the graduation project in

Artificial Intelligence and Data Science

Prepared by :

Hamza Zaher Alsamman

Raghad Mohamed Ali Abdulrahman

Supervisors :

Dr. Majida Albakoor

Eng. Wesam Alsohle

Academic Year 2025/2026