الجامعة السورية الخاصة

كلية هندسة الذكاء الصنعي

قسم هندسة الذكاء الصنعي وعلوم البيانات الذكية

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

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

أعدت هذه الأطروحة

لإنجاز المشروع الفصلي في اختصاص الذكاء الصنعي وعلوم البيانات

اعداد الطلاب :

حمزة زاهر السمان

رغد محمد علي عبد الرحمن

أسماء المشرفين :

د.ماجدة البكور

م.آية الأسود

العام الدراسي 2026/2025

الملخص :

يتناول هذا المشروع تصميم وتنفيذ نظام دبلجة آلي متكامل للفيديوهات التعليمية، يهدف إلى تعريب المحتوى 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 تتجاوز عقبات التكلفة والزمن في الطرق التقليدية.

تكمن الإضافة النوعية لهذا البحث في **بناء وتدريب نموذج شبكة عصبونية عميقة تعتمد على بنية الـU 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 تقنيات **WhisperX** المتقدمة لتحقيق التعرف الآلي على الكلام والترجمة، مع الاعتماد على ميزة المحاذاة الزمنية الدقيقة (Phoneme-level Alignment) لضمان مطابقة الصوت المترجم مع الجدول الزمني للفيديو الأصلي بدقة متناهية.

أظهرت النتائج العملية كفاءة عالية للنظام في تحسين جودة الصوت وتقليل الأخطاء الزمنية؛ حيث أثبتت التجارب قدرة المنظومة على معالجة ودبلجة فيديو تعليمي مدته **8 دقائق** في زمن **12 دقائق** باستخدام معالج رسوميGPU ، مما يجعله حلاً فعالاً وقابلاً للتوسع في تعريب منصات التعليم المفتوح

Abstract :

system for educational videos, translating content from English to Arabic using deep learning techniques. The research addresses the linguistic communication gap hindering access to global academic content and the necessity for rapid dubbing systems that overcome the cost and time constraints of traditional methods.

The core contribution of this work lies in **building and training a deep neural network model based on the U-Net architecture** specifically for speech-noise separation and enhancement. By developing and training this architecture, the system ensures high-quality audio extraction from educational videos before linguistic processing, distinguishing this work from existing studies that often rely on pre-built models. This custom model is integrated with advanced **WhisperX** technology for Automatic Speech Recognition (ASR) and translation, utilizing phoneme-level alignment to ensure precise synchronization between the translated audio and the original video timeline.

Experimental results demonstrate high system efficiency in audio quality improvement and temporal accuracy. Tests show that the system can process and dub a **8-minutes** educational video in approximately **12 minutes** using GPU acceleration, providing an effective and scalable solution for localizing open educational platforms.

الفهرس

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

الفصل الأول
المقدمة

الفصل الاول - المقدمة

1.1 تمهيد (Introduction)

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 تحويل الفيديوهات التعليمية من الإنجليزية إلى العربية آلياً، مع ضمان دقة المحتوى وجعل التعلم المفتوح متاحاً للجمهور العربي بكفاءة وسرعة.

2.1 مشكلة البحث (Problem Statement)

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 مطابقة النص المترجم مع الطوابع الزمنية للفيديو الأصلي بشكل دقيق (دقة التزامن). كما تبرز مشكلة التكلفة والزمن، حيث تتطلب الدبلجة البشرية موارد مالية ضخمة ووقتاً طويلاً للمعالجة. يعالج هذا البحث هذه المشكلات عبر بناء منظومة متكاملة تبدأ بموديل **U-Net** مُدرب خصيصاً لفصل الضجيج، متبوعاً بنظام **WhisperX** لضمان الترجمة الدقيقة والمحاذاة الزمنية الاحترافية.

3.1 أهداف البحث (Research Objectives)

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 لتطابق حركة الشفاه والفيديو الأصلي باستخدام تقنيات المحاذاة على مستوى "الفونيم."
3. تقليل الفجوة اللغوية في الوصول للمعرفة العالمية عبر توفير حل دبلجة سريع ومنخفض التكلفة .
4. بناء نظام ذو كفاءة زمنية عالية يتيح معالجة الفيديوهات الطويلة في زمن يقارب الوقت الحقيقي

(Real-time processing)

4.1النتائج والإسهامات (Contributions and Results)

قدم البحث حلولاً تقنية مبتكرة في مجال المعالجة الصوتية واللغوية، حيث تم بناء وتدريب نموذج **U-Net** عصبوني عميق مخصص لفصل الضجيج وتعزيز الكلام، مما ساهم بفاعلية في تحسين نقاء الإشارة الصوتية قبل البدء بعملية الترجمة. كما تم تحقيق محاذاة زمنية فائقة الدقة عبر دمج تقنيات **WhisperX** التي تتيح محاذاة النص المترجم على مستوى **"الفونيم "**(وهو أصغر وحدة صوتية في الكلام)؛ مما يضمن تزامن الدبلجة مع حركة الفيديو الأصلي بدقة متناهية تتجاوز كفاءة الأنظمة التقليدية. بالإضافة إلى ذلك، أظهر النظام كفاءة عالية في الأداء الزمني، حيث أثبتت التجارب القدرة على معالجة ودبلجة فيديو تعليمي مدته 8 دقائق خلال 12 دقائق فقط باستخدام معالجاتGPU ، مما يعزز من إمكانيات استخدامه في تطبيقات تعريب المحتوى التعليمي الضخم بشكل آلي وسريع.

5.1 هيكلية البحث (Research Structure)

يتألف هذا البحث من ستة فصول مترابطة تتبع الهيكلية الأعتمدة في الجامعة؛ حيث يتناول الفصل الأول المقدمة وأهمية البحث، بينما يستعرض الفصل الثاني الأدبيات السابقة حول أنظمة معالجة الصوت والتعرف على الكلام. ويركز الفصل الثالث على المنهجية المستخدمة بدءاً من تصميم نموذج U-Net وصولاً إلى تطبيق .WhisperX 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(Fundamental Concepts)

قبل الخوض في التفاصيل التقنية، لا بد من تأطير 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 وحس الواقعية في الفيديو.

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 (AI Dubbing):** وهي الثورة الحالية التي مكنت من أتمتة العملية بالكامل عبر تكامل نماذج التعرف على الكلام (ASR) ، والترجمة الآلية (MT) ، وتوليد الكلام (TTS) ، مما وفر حلولاً سريعة ومنخفضة التكلفة خاصة للمحتوى التعليمي الضخم.

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. **المحاذاة الزمنية الدقيقة :**استخدام تقنيات المحاذاة القسرية (Forced Alignment) لضمان أن كل كلمة تنطق في وقتها الصحيح تماماً.
3. **واقعية النطق :**استخدام نماذج توليد الكلام (TTS) التي تحاكي النبرة البشرية الطبيعية.

تنوعت الجهود البحثية في بناء أنظمة الدبلجة الآلية ومعالجة الكلام عبر تكامل تقنيات التعرف على الكلام، عزل الصوت، والمزامنة البصرية. ففي سياق التعرف الدقيق على الكلام، طورت**[دراسة 1]** نظام **WhisperX**، والذي يمثل تطوراً جوهرياً لنموذج Whisper الأصلي؛ حيث اعتمدت الدراسة على استراتيجية "القطع والدمج (Cut & Merge) "المعتمدة على كاشف النشاط الصوتي (VAD) ونموذج المحاذاة القسرية للفونيمات (Forced Phoneme Alignment) لضمان توقيتات دقيقة على مستوى الكلمة، وهو ما عالج مشكلات "الهلوسة" والانزياح الزمني، محققة تسارعاً في التنفيذ يصل إلى 12 ضعفاً.

وعلى صعيد تحديد هوية المتحدثين، استعرضت **[دراسة 2]**نظام **pyannote.audio 2.1** كمنهجية متقدمة لتقسيم وتحديد المتحدثين في البيئات المعقدة؛ إلا أن مشروعنا ركّز على سيناريو **المتحدث الواحد(Single Speaker)** ، مما سمح بتبسيط هذه المرحلة وتوجيه موارد المعالجة لضمان استقرار الهوية الصوتية للمتحدث الوحيد دون الحاجة لتعقيدات الفصل بين هويات متعددة.

وفيما يخص جودة المدخلات الصوتية، تناولت **[دراسة 3]**تقنيات عزل المصادر مثل نظام **VAST**الذي يعتمد على الإشراف الذاتي. وفي هذا السياق، وبدلاً من الاعتماد على مكتبات التعرف على الهوية في التنقية، قمنا ببناء **نموذج عزل الضوضاء (Speech-Noise Separation)**الخاص لتعزيز جودة الإشارة وتنقيتها من المؤثرات الخارجية قبل مرحلة الدبلجة، مما يضمن نقاء صوت المتحدث الوحيد ووضوحه قبل المعالجة اللاحقة.

أما في مرحلة تخليق الكلام، فقد ركزت التوجهات الحديثة مثل **[دراسة 4]**على نموذج **XTTS**، الذي حقق طفرة في استنساخ البصمة الصوتية (Voice Cloning) بدقة عالية عبر لغات متعددة. ومع ذلك، اعتمد مشروعنا في التنفيذ العملي على تقنية **Edge TTS**؛ وذلك لعدة أسباب استراتيجية أهمها: تفوقها في جودة ونقاء المخارج الصوتية للغة العربية، وسرعة الاستجابة الاستثنائية (Low Latency) بفضل المعالجة السحابية، بالإضافة إلى توفير نبرات صوتية طبيعية (Prosody) تتلاءم مع سياق الحوار دون الحاجة لمتطلبات عتادية ضخمة كما هو الحال في نماذج XTTS ، مما يضمن توازناً مثالياً بين الدقة اللغوية وكفاءة التشغيل.

وختاما تستعرض **[دراسة 5]** آفاق المزامنة البصرية عبر نموذج **Wav2Lip**، الذي يعتمد على 'مميز خبير (Expert Discriminator)' تم تدريبه مسبقاً لاكتشاف عدم التطابق بين الإشارة الصوتية وحركة الشفاه. ورغم أن هذا البحث يركز بشكل أساسي على المحاذاة الصوتية واللغوية، إلا أن تقنية Wav2Lip تمثل **المعيار التقني المستهدف** لضمان تطابق حركة شفاه المتحدث مع النص العربي المولد في مراحل التطوير المتقدمة، حيث أثبتت الدراسة قدرة النموذج على العمل مع أي وجه وأي لغة بدقة تقارب الواقع.

تؤكد هذه الأدبيات فاعلية البنيات الهجينة التي تجمع بين دقة التوقيت في WhisperX، وجودة التوليد الصوتي في Edge **TTS** يبرز مشروعنا كنموذج تطبيقي يركز على بناء **النواة الصوتية واللغوية للدبلجة** مخصصة للمتحدث الواحد، مع معالجة تحديات الضجيج عبر نموذج U-Net مستقل. وبذلك، يضع هذا البحث حجر الأساس الذي يمكن البناء عليه مستقبلاً لدمج تقنيات المزامنة البصرية مثل **Wav2Lip** ، لتقديم تجربة دبلجة عربية كاملة (صوتياً وبصرياً).

ويمكن تلخيص الدراسات المرجعية في هذا الجدول:

جدول(1-2):

مقارنة تقنية بين منهجيات ومعايير قياس الأبحاث السابقة

الدراسة	آلية العمل (Methodology)	معايير التقييم (Metrics)
WhisperX [1]	يعتمد استراتيجية **Forced Phoneme Alignment** عبر دمج نموذج Whisper مع نماذج فونيمية (مثل wav2vec2) لربط النص بالصوت بدقة زمنية عالية.	**WER** (نسبة خطأ الكلمات)، **RTF** (عامل الوقت الحقيقي)، ودقة الطوابع الزمنية.
[2] pyannote.audio	بنية عصبونية تعتمد على **EEND-VC** لتقسيم المتحدثين، وتستخدم تقنيات العنقودة (Clustering) لتحديد هوية كل متحدث في المقطع.	**DER** (معدل خطأ التقسيم)، و **JER** (معدل خطأ جاكارد) لتقييم دقة تحديد هوية المتحدث.
VAST [3]	يعتمد على **Trimodal Consistency** (الاتساق الثلاثي) بين الصوت، اللغة، والرؤية عبر تقنية التعلم التباينيي (Contrastive Learning) لعزل المصادر.	**SDR** (نسبة الإشارة إلى التشوه)، و **CLAPScore** لتقييم الصلة الدلالية بين النص والصوت المعزول.
XTTS [4]	نموذج لغوي (GPT-based) يعتمد على **Conditioning Encoder** لاستخلاص بصمة المتحدث وتوليد كلام بأسلوب "Zero-shot" عبر لغات متعددة.	**CER** (نسبة خطأ الحروف)، **SECS** (تشابه البصمة الصوتية)، و **UTMOS** للتقييم الذاتي للجودة.
Wav2Lip [5]	يستخدم شبكة **GAN** مع مميز خبير (Expert Discriminator) مدرب مسبقاً (SyncNet) لفرض دقة حركة الشفاه وتوافقها مع الكلام المولد.	**LSE-D** (مسافة خطأ المزامنة)، **LSE-C** (ثقة المزامنة)، و **FID** لتقييم جودة الصورة.

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 (Audio Source Separation):

- **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 استدعاء هذا الموديل لعزل صوت الموسيقى تماماً عن صوت الكلام، مما يضمن الحصول على مسار بشري نقي قبل البدء بالترجمة.

3.1.2 نموذج ResUNet المطور (الموديل الخاص بنا):

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

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

3.1.3تقنية WhisperX للمحاذاة والتعرف على الكلام:

- **ما هوWhisperX** ؟ هو نسخة "محسنة" ومطورة من نموذج Whisper الأصلي. هو ليس مجرد أداة لتحويل الصوت إلى نص، بل هو نظام متكامل للمحاذاة الزمنية الدقيقة.(Time Alignment)

- **دوره في المشروع :**هو المسؤول عن "فهم" ما يقال وتحويله إلى نص مترجم، والأهم من ذلك هو تحديد **متى قيلت كل كلمة بالضبط**.

- **وظيفته :**يوفر "طوابع زمنية (Timestamps) "على مستوى الكلمة وحتى مستوى "الفونيم" (أصغر وحدة صوتية)، مما يسمح للنظام بوضع الصوت العربي المدبلج في مكانه الصحيح تماماً ليتطابق مع حركة فم المحاضر.

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

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

جدول(1-3)

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 فونيمية** 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.

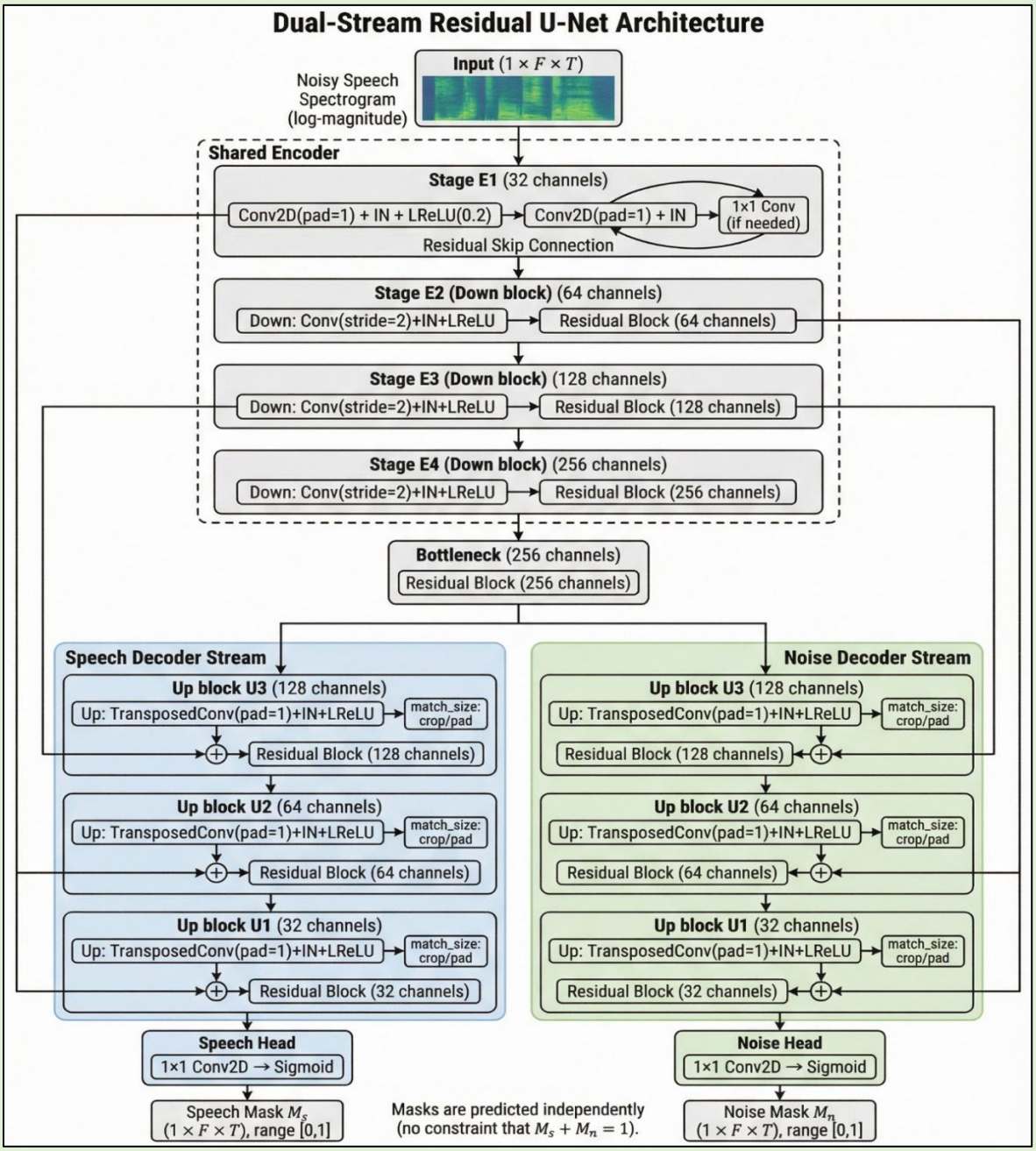
3.2 التحليل التقني لخوارزمية الشبكة العصبونية التلافيفية (U-Net/CNN):

3.2.1 مقدمة حول الشبكة التلافيفية:

تعتبر الشبكات التلافيفية (Convolutional Neural Networks - CNN) وبنيتها المتطورة **U-Net** الأنسب لمعالجة الإشارات الصوتية بعد تحويلها لمخططات طيفية؛ حيث تعتمد فكرة العمل على استخدام طبقات تلافيفية (Convolutional Layers) لاستخراج الميزات الهرمية من الترددات الصوتية، مما يجعلها فعالة للغاية في فصل الكلام عن الضجيج في المحاضرات التعليمية.

3.2.2 بنية النموذج وتقسيم الطبقات المستخدمة:

يعتمد النموذج على بنية مشفر وفك تشفير (Encoder-Decoder) من نوع ResUNet ، مصممة لتحقيق توازن مثالي بين استعادة جودة الصوت وسرعة المعالجة .ويوضح **الشكل (3-1)** 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.



شكل (1-3): بنية الشبكة العصبونية

(Dual-Stream Residual U-Net)

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

3.2.3 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 "الفونيم"، 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 الرابع
التجارب والنتائج والتقييم

4.1 التحديات والاحتياجات (Challenges and Requirements): قبل البدء بتنفيذ النظام، واجهنا مجموعة من التحديات الجوهرية التي تطلبت استراتيجيات معالجة خاصة:

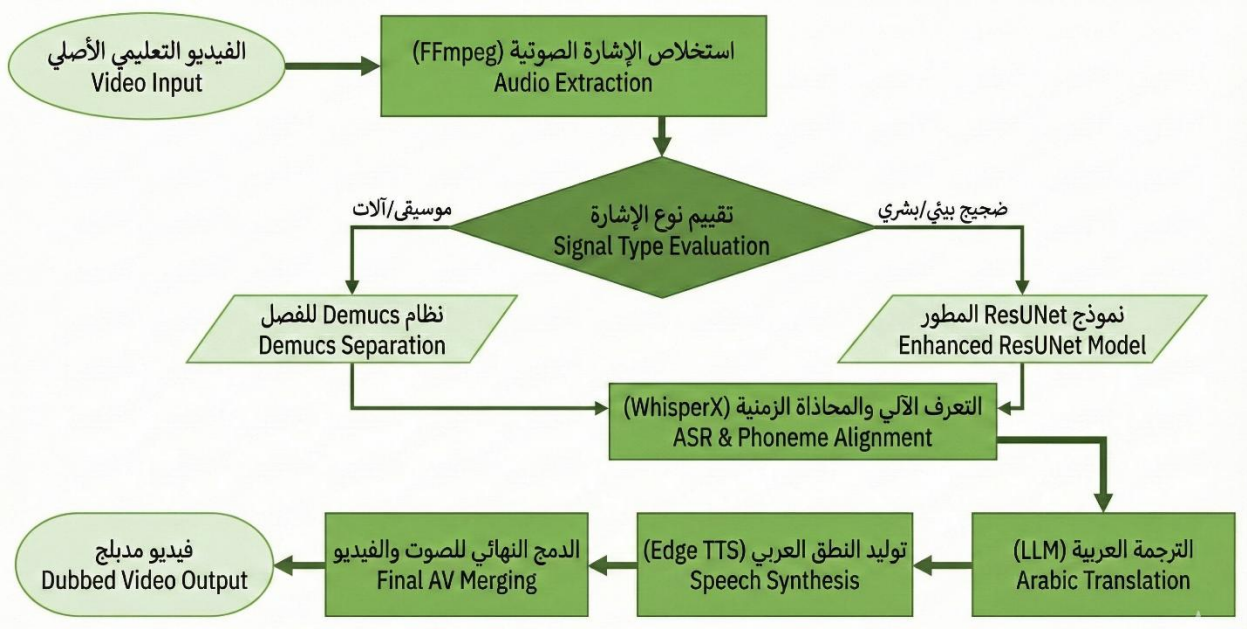
- **مجموعات المعطيات:** تطلبت الشبكات التلافيفية بيانات ضخمة ومتنوعة للتكيف مع أنماط الضجيج المختلفة.
- **تعقيد الإشارة:** تداخل بعض الأصوات التعليمية مع موسيقى خلفية أو صدى، مما تطلب دقة عالية في استخلاص السمات.
- **التفاعل مع البيئة:** الحاجة لتطوير قدرات الموديل ليتكيف مع اختلاف جودة الميكروفونات في التسجيلات التعليمية.

4.2 مقاييس التقييم (Evaluation Metrics): للحكم على جودة النظام وفصل الصوت، تم اعتماد مجموعة من المقاييس الرياضية المعيارية. يوضح **الجدول (1-4)** المقاييس المستخدمة والهدف من كل منها، حيث تعطي هذه المقاييس مؤشراً دقيقاً عن مدى وضوح الكلام ونسبة الخطأ اللغوي

جدول (1-4): 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.

4.3 مخطط المنهجية العام (Methodology Framework): يوضح **الشكل (1-4)** المنهجية المتسلسلة المتبعة في النظام، حيث نلاحظ أن العملية تبدأ بتحليل نوع الإشارة الصوتية لضمان عزل دقيق للضجيج عبر مسار (ResUNet) أو (Demucs).تلي ذلك مرحلة التعرف والمحاذاة باستخدامWhisperX ، ثم تأتي الخطوة الجوهرية وهي **الترجمة العربية (Arabic Translation)**التي تحول النص المستخلص إلى لغة الهدف، ليتم بعد ذلك تمرير النص المترجم إلى محرك **توليد النطق العربي (Edge TTS)** لتحويله إلى إشارة صوتية طبيعية، وتنتهي المنهجية بدمج المسار الصوتي الجديد مع الفيديو الأصلي لضمان تزامن احترافي.



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

4.4 مجموعة المعطيات المستخدمة وتحضيرها (Datasets): تم الاعتماد على مجموعة بيانات-VoiceBank DEMAND، وهي مزيج بين VCTK للكلام النقي و DEMAND للضجيج البيئي. يوضح **الجدول (2-4)** إحصائيات البيانات وكيفية تقسيمها لضمان عدم حدوث تسريب للبيانات (Data Leakage) عبر عزل متحدثين محددين للاختبار فقط.

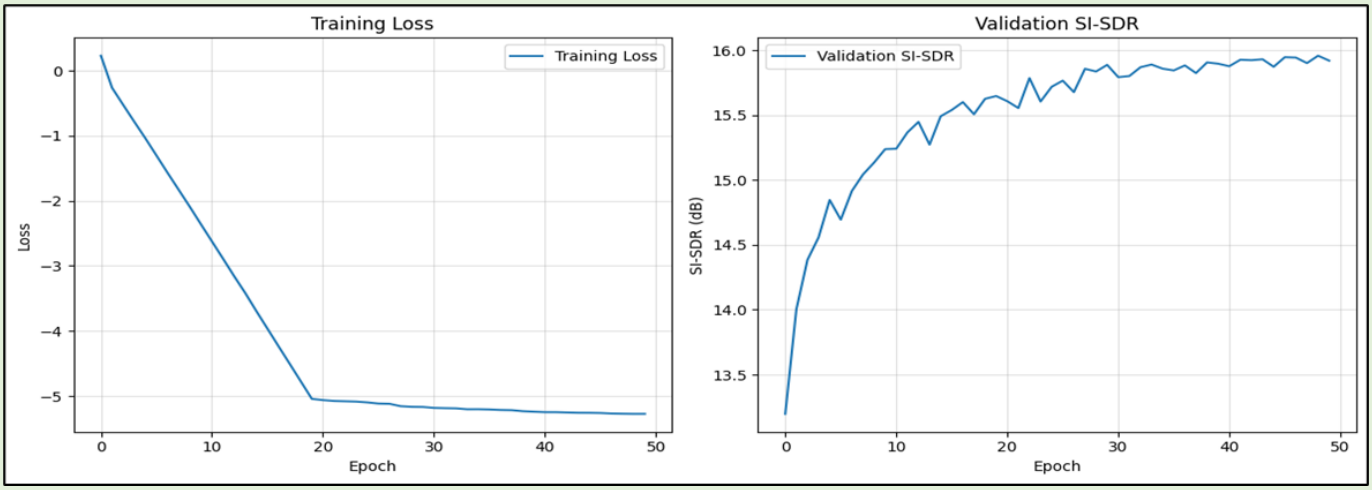
جدول (2-4): إحصائيات تقسيم مجموعة البيانات المستخدمة

المعيار	التفاصيل الإحصائية
معدل العينة	16,000 هرتز (16 kHz) لضمان استقرار النماذج.
توزيع التدريب	24 متحدثاً (حوالي 85% من إجمالي البيانات).
توزيع الاختبار	4 متحدثين معزولين تماماً (حوالي 15%) لضمان قدرة النموذج على التعميم.

4.5 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)

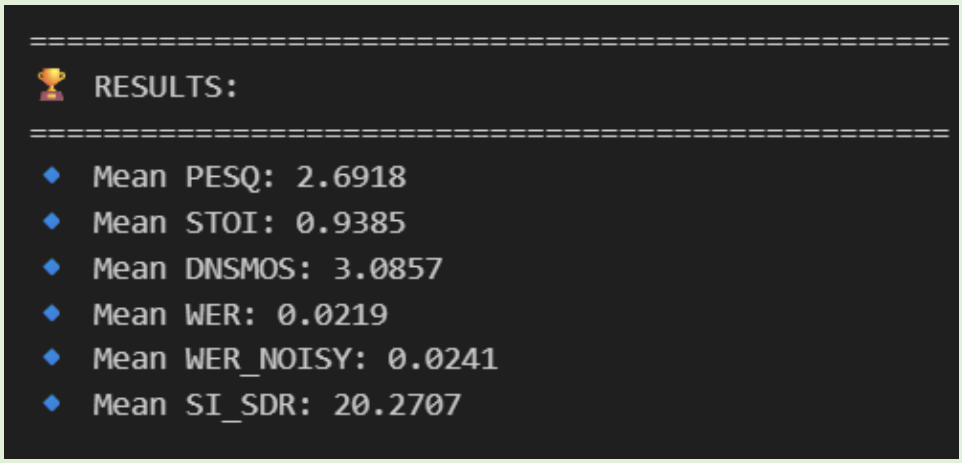
ويوضح **الشكل (2-4)** منحنيات الأداء أثناء عملية التدريب؛ حيث نلاحظ الانخفاض التدريجي والمستقر في دالة الخسارة (Training Loss)، بالتوازي مع الارتفاع الملحوظ في مقياس (SI-SDR) لمجموعة التحقق، مما يثبت وصول النموذج إلى حالة التقارب (Convergence) المنشودة دون حدوث إفراط في التدريب (Overfitting) خلال الـ 50 حقبة المحددة.



شكل (2-4): منحنيات تطور دالة الخسارة ومقياس SI-SDR أثناء تدريب النموذج

4.6 نتائج الاختبار وتحليل الأداء: بعد إتمام مراحل التدريب والمعالجة، تم تقييم أداء النظام باستخدام عينة الاختبار من مجموعة بيانات .VoiceBank-DEMAND يوضح **الشكل (2-4)** 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.

الشكل (2-4): النتائج الإحصائية النهائية لأداء نظام الدبلجة الآلي



الفصل الخامس
الخاتمة والتوصيات المستقبلية

الفصل الخامس -الخاتمة والآفاق المستقبلية

1.5 الخاتمة:

تم في هذا البحث استخدام تقنيات تعلم عميق متقدمة لتطوير نموذج فعال لدبلجة الفيديوهات التعليمية من اللغة الإنجليزية إلى العربية، وذلك عبر دمج معمارية **U-Net**لمعالجة وفصل الضجيج الصوتي، وتقنيات **WhisperX** للمحاذاة الزمنية الدقيقة أظهرت التجارب أن النموذج يحقق أداءً متزناً في تحسين جودة الإشارة الصوتية ودقة التعرف على الكلام، مع قدرة عالية على التعامل مع السيناريوهات التعليمية المختلفة التي تشوبها ضوضاء الخلفية.

بلغت دقة النموذج في التعرف على الكلام (**WER**) نسبة خطأ ضئيلة جداً وصلت إلى **0.02**، بينما حقق مقياس جودة الصوت (**PESQ**) قيمة **2.69** على مجموعة بيانات الاختبار، مما يشير إلى نجاح كبير في الحفاظ على مفهومية النطق البشري. وبالرغم من ذلك، لا يزال هناك مجال واسع للتحسين في جوانب مثل محاكاة نبرة الصوت الأصلية (Voice Cloning) والتعامل مع التداخلات الصوتية المعقدة في الفيديوهات ذات الجودة المتدنية جداً.

تم تقييم النظام باستخدام معايير أساسية شملت **STOI** و **DNSMOS**، والتي أظهرت أن المنظومة قادرة على تحقيق نتائج مرضية جداً مع التوصية ببعض التحسينات لتعزيز المتانة في البيئات الواقعية. يمكن لهذا النموذج أن يشكل حجر زاوية لتطبيقات عملية في تعريب المنصات الأكاديمية العالمية، ودعم أنظمة التعليم المفتوح، وأتمتة إنتاج المحتوى المرئي المترجم.

2.5 الآفاق المستقبلية:

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

1. **تحسين جودة استنساخ الصوت (Voice Cloning):**
   - تطوير تقنيات متقدمة تتيح للنظام دبلجة الفيديو بنفس نبرة صوت المحاضر الأصلي، مما يساهم في تحسين تجربة المستخدم وجعل الدبلجة تبدو طبيعية وأكثر واقعية.
2. **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.
3. **استخدام تقنيات تعلم عميق هجينة:**
   - تجربة دمج بنى شبكات عصبونية أكثر تطوراً مثل **Transformers** لمعالجة الصوت أو شبكات **GANs** لتحقيق مزامنة بصرية (Lip-Sync) بين حركة الشفاه والصوت العربي الناتج، لتعزيز الجوانب الزمنية والجمالية للفيديو.
   - تطبيق تقنيات التعلم المستمر (**Continual Learning**) لتكييف النموذج مع البيانات الجديدة دون الحاجة لإعادة التدريب الكامل، مما يعزز قدرته على التطور مع مرور الوقت.
4. **تحسين عمليات المعالجة الفورية (Streaming):**
   - تطوير تقنيات لمعالجة التدفقات الصوتية مباشرة، مما يسمح بالترجمة والدبلجة الفورية أثناء البث المباشر للمحاضرات والندوات التعليمية بدقة تفوق الأداء الحالي.
5. **جمع أنواع مختلفة من المعطيات (Multimodal):**
   - استغلال البيانات البصرية (حركة الفم والوجه) لتعزيز دقة التعرف على الكلام في حالات الضجيج الشديد، مما يحسن سياق التعرف والترجمة.

3.5 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.

4.5 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] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020, October). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 484-492).

study link

Syrian Private University

Faculty of Engineering

Artificial intelligence and Data science

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

A Thesis Prepared for the Fulfillment of the Requirements for the Semester Project in the Artificial Intelligence and Data Science Department

Prepared by:

Hamza Zaher Alsamman

Raghad Mohamed Ali Abdulrahman

Supervisors:

Dr. Majida Albakoor

Eng. Aya Alaswad

Academic Year 2025/2026