الجامعة السورية الخاصة

كلية الهندسة

قسم الذكاء الصنعي وعلوم البيانات

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 :

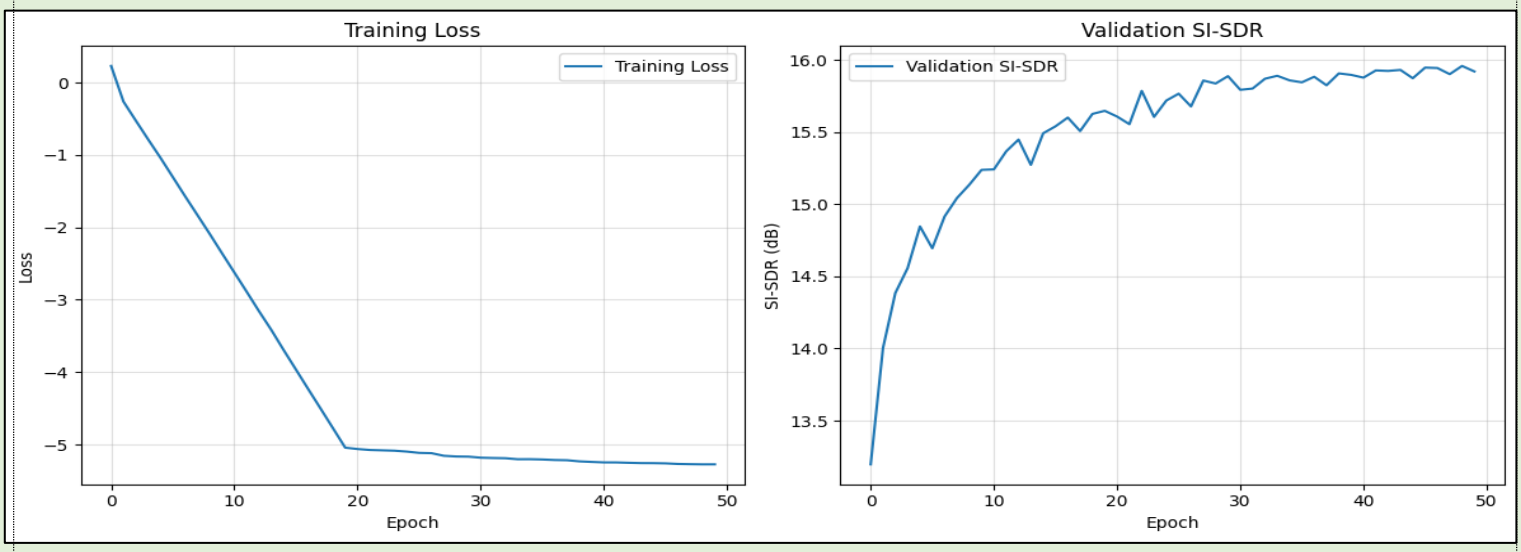
system for educational videos, translating content from English to Arabic using deep learning techniques. The research addresses the linguistic communication gap hindering access to global academic content and the necessity for rapid dubbing systems that overcome the cost and time constraints of traditional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

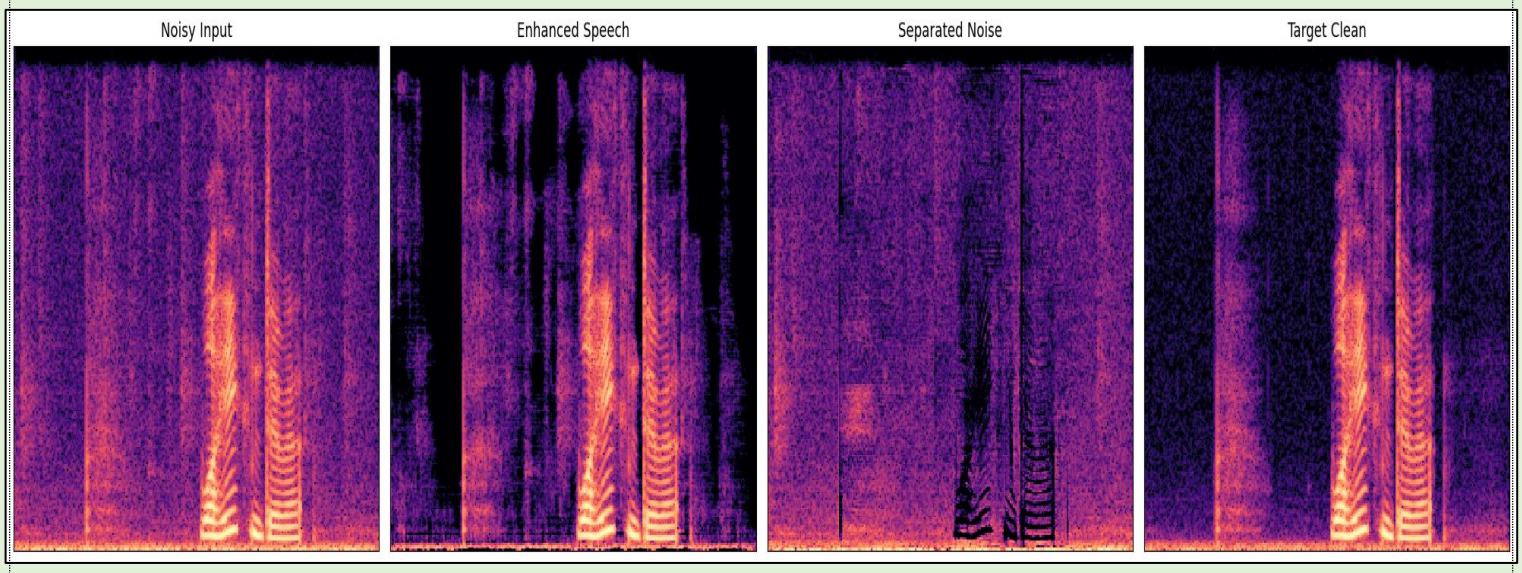
Experimental results demonstrate high system efficiency in audio quality improvement and temporal accuracy. Tests show that the system can process and dub a **10-minute** educational video in approximately **4 minutes** using GPU acceleration, providing an effective and scalable solution for localizing open educational platforms.

الفهرس

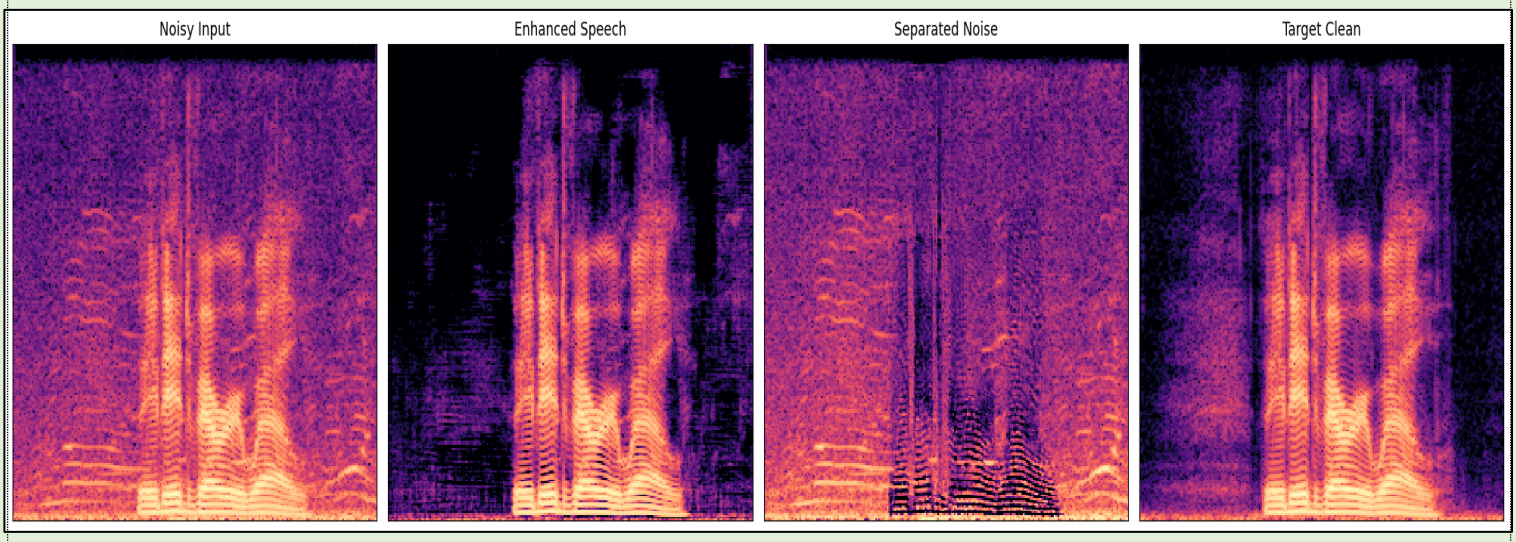
فهرس الأشكال



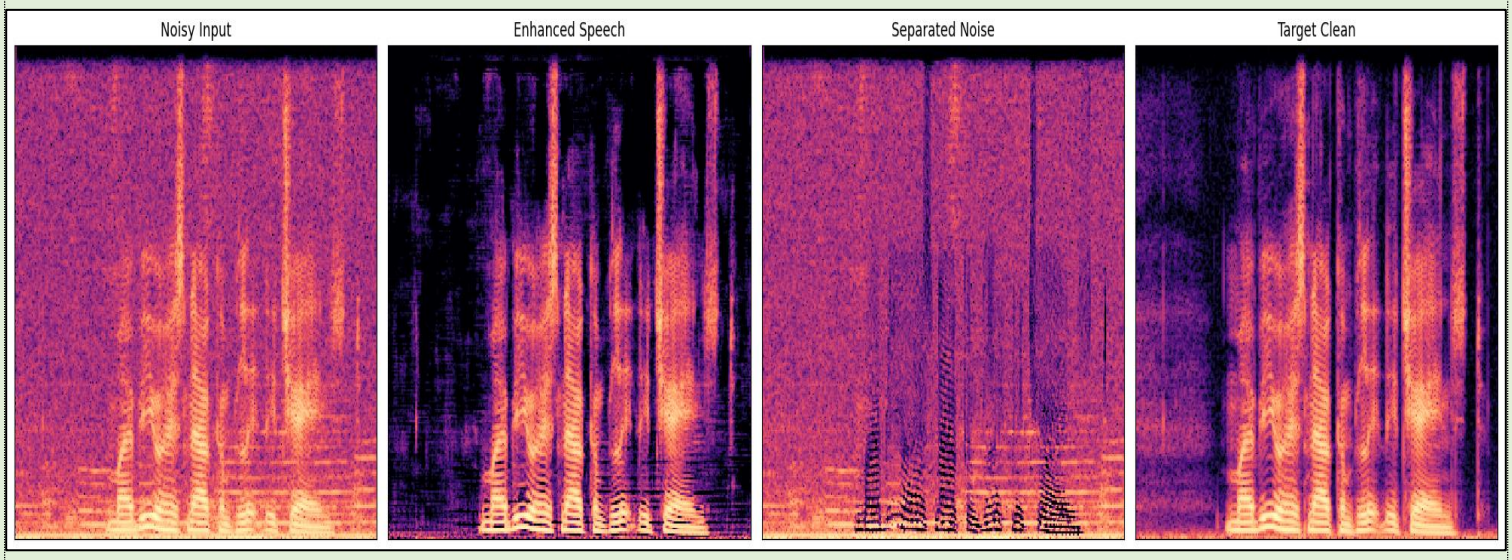
الشكل (1)



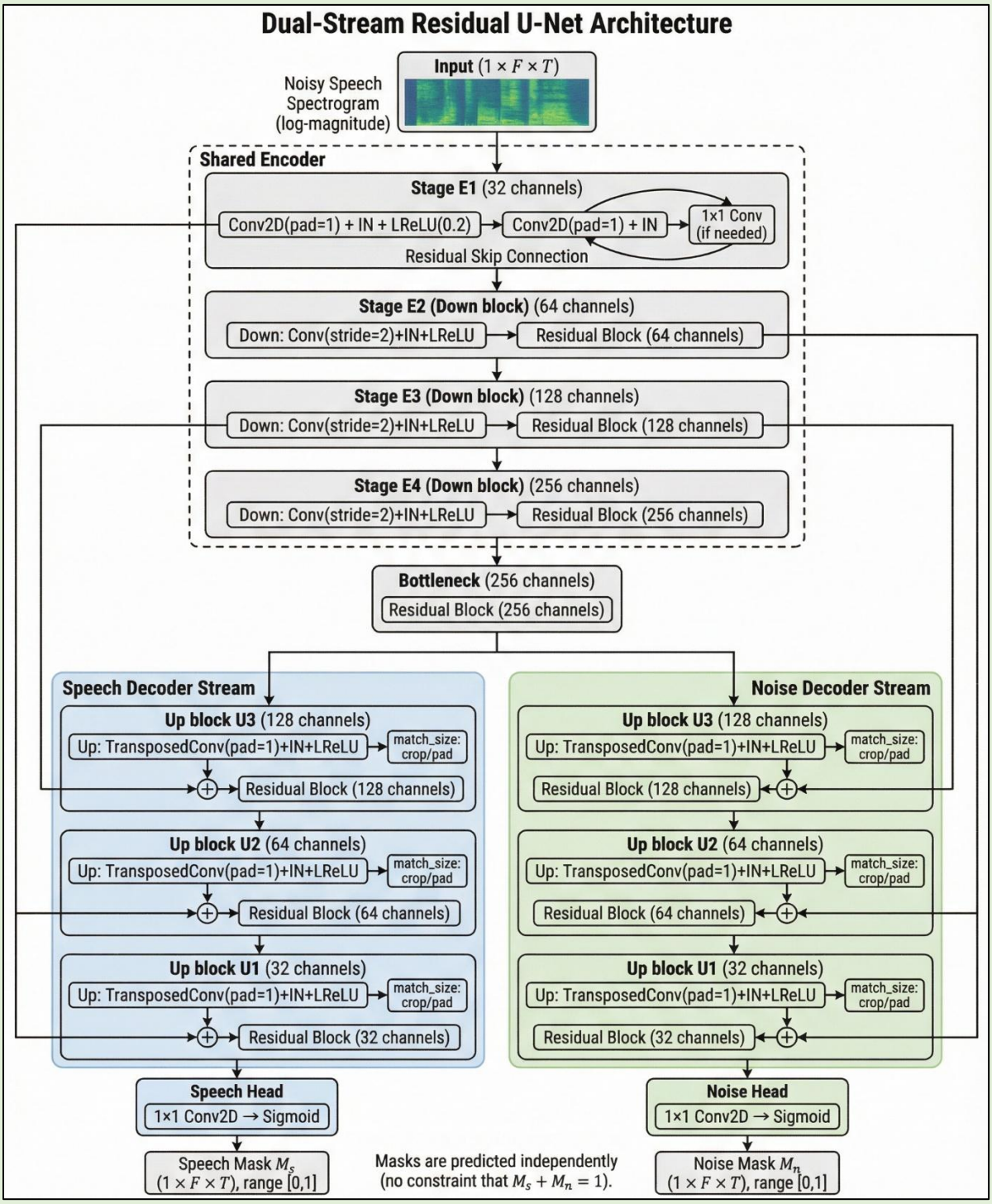
الشكل (2)



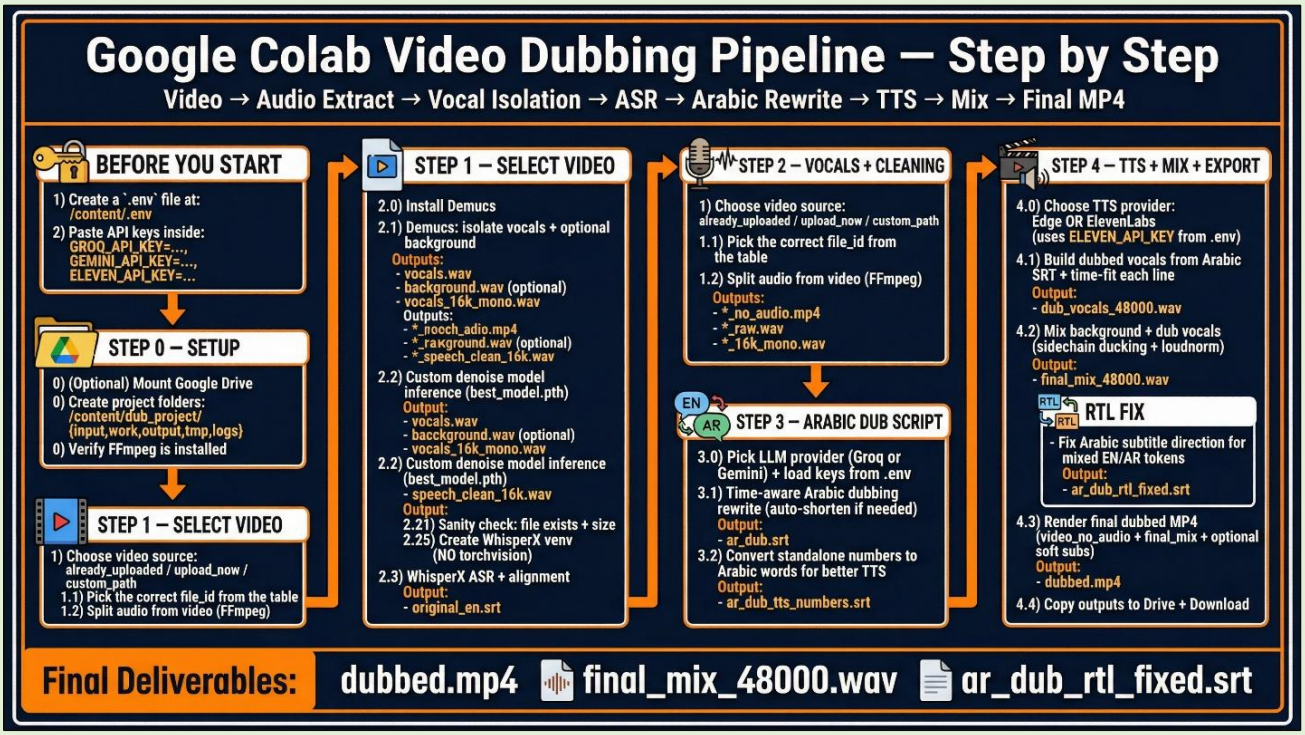
الشكل (3)



الشكل (4)


Dual-Stream Residual U-Net Architecture
Input ($1 \times F \times T$)
Noisy Speech Spectrogram (log-magnitude)
Shared Encoder
Stage E1 (32 channels)
Conv2D(pad=1) + IN + LReLU(0.2)
Conv2D(pad=1) + IN
1×1 Conv (if needed)
Residual Skip Connection
Stage E2 (Down block) (64 channels)
Down: Conv(stride=2)+IN+LReLU
Residual Block (64 channels)
Stage E3 (Down block) (128 channels)
Down: Conv(stride=2)+IN+LReLU
Residual Block (128 channels)
Stage E4 (Down block) (256 channels)
Down: Conv(stride=2)+IN+LReLU
Residual Block (256 channels)
Bottleneck (256 channels)
Residual Block (256 channels)
Speech Decoder Stream
Noise Decoder Stream
Up block U3 (128 channels)
Up: TransposedConv(pad=1)+IN+LReLU
match_size: crop/pad
Residual Block (128 channels)
Up block U2 (64 channels)
Up: TransposedConv(pad=1)+IN+LReLU
match_size: crop/pad
Residual Block (64 channels)
Up block U1 (32 channels)
Up: TransposedConv(pad=1)+IN+LReLU
match_size: crop/pad
Residual Block (32 channels)
Speech Head
1×1 Conv2D → Sigmoid
Noise Head
1×1 Conv2D → Sigmoid
Speech Mask M_s ($1 \times F \times T$), range [0,1]
Masks are predicted independently (no constraint that $M_s + M_n = 1$).
Noise Mask M_n ($1 \times F \times T$), range [0,1]

الشكل (5)



الشكل (6)

فهرس النتائج

```
==================================================
🏆 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
```

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 النتائج 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.

4.1 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
الأدبيات السابقة

الفصل الثاني – الأدبيات السابقة

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).

الفصل الثالث
المفاهيم الأساسية والخصائص

الفصل الثالث – المنهجية

1.3المنهجية - المفاهيم الأساسية والخصائص

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

1.1.3 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.

2.1.3نموذج 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.

4.1.3المقارنة بين Whisper (الأصلي) و WhisperX (المطور):

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

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.

2.3التحليل التقني لخوارزمية الشبكة العصبونية التلافيفية (U-Net/CNN):

1.2.3مقدمة 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.

2.2.3بنية 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، وتتكون 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

3.2.3التحسينات 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.

3.3التحديات والاحتياجات:

- **مجموعات المعطيات:** تتطلب الشبكات التلافيفية بيانات ضخمة ومتنوعة (LibriSpeech + UrbanSound) للتكيف مع أنماط الضجيج المختلفة.
- **تعقيد الإشارة:** بعض الأصوات التعليمية تتداخل مع موسيقى خلفية أو صدى، مما يتطلب دقة عالية في استخلاص السمات.
- **التفاعل مع البيئة:** يجب تطوير قدرات الموديل ليتكيف مع اختلاف جودة الميكروفونات في التسجيلات التعليمية المتنوعة.

4.3مقاييس التقييم:

تم استخدام عدة مقاييس لتحديد جودة النظام وفصل الصوت، وأبرزها:

- **PESQ:** لتقييم جودة الصوت الإدراكية (حقق النظام 2.69).
- **STOI:** لقياس مدى وضوح وفهم الكلام (حقق النظام 0.93).
- **WER (Word Error Rate):** لقياس دقة التعرف على الكلام (حقق نسبة خطأ منخفضة جداً بلغت 0.02).
- **DNSMOS:** مقياس موضوعي لجودة فصل الضجيج (حقق 3.08).

5.3الاستراتيجيات والتقنيات التنفيذية:

- **المعالجة عبر FFmpeg:** لفصل الصوت وإعادة العينة لتردد 16 kHz لضمان استقرار النماذج.
- **المحسن Adam:** لضبط الأوزان وتحسين جودة التدريب.
- **زيادة البيانات(Data Augmentation):** عبر دمج الضجيج عشوائياً لتعزيز متانة النموذج.
- **المعالجة المتوازية (GPU):** استخدام Kaggle لتسريع التدريب والاستدلال.

6.3الفجوات المعرفية والتطبيقية وكيفية معالجتها:

- **تنوع المعطيات:** تم استخدام مجموعة بيانات (VoiceBank-DEMAND 16kHz) لتدريب النموذج على بيئات ضجيج واقعية ومختلفة.
- **استقرار الأداء:** بدلاً من التوقف المفاجئ، تم استخدام تقنية **Learning Rate Warmup** في البداية لضمان استقرار الأوزان، متبوعة بجدولة ذكية لمعدل التعلم تمنع حدوث التذبذب في قيم الخسارة (Loss)

7.3منهجية العمل:

تعتمد المنهجية المتبعة في هذا المشروع على أسلوب المعالجة الهجينة وتدفق البيانات الذكي(Hybrid Pipeline) ؛ حيث تبدأ العملية بتقييم الإشارة الصوتية لتحديد مسار الفصل الأنسب: يتم توجيه المقاطع التي تحتوي على موسيقى خلفية أو آلات لنظام **Demucs**، بينما يتم معالجة المقاطع التي تحتوي على ضجيج بشري أو بيئي بواسطة نموذج **ResUNet**المطور ضمن هذا البحث. تلي ذلك مرحلة المحاذاة والمزامنة باستخدام تقنية **WhisperX**لضمان دقة الدبلجة.

- **عدد الحقب (50 Epochs):** تم تحديدها لضمان وصول النموذج إلى حالة التقارب الكامل (Convergence) واستيعاب الفرق بين أنماط الضجيج المختلفة.
- **معدل التعلم والجدولة:** بدأ التدريب بمعدل **1e-4**مع استخدام جدولة **ReduceLROnPlateau**، التي تقوم بخفض المعدل تلقائياً عند ثبات الأداء لضمان الاستقرار في النتائج النهائية وتجنب التذبذب في قيم الخسارة (Loss)
- **حجم الدفعة :(Batch Size = 16)** تم اختيارها لتناسب سعة ذاكرة معالجات الرسوميات (GPU) المستخدمة في بيئة **Kaggle** مثل(Tesla T4) ، ولضمان استقرار حسابات التدرج (Gradient Calculations) أثناء تدريب نموذج ResUNet.

8.3مجموعة المعطيات المستخدمة وتحضيرها:

تعتمد كفاءة النموذج على جودة البيانات ومدى محاكاتها للواقع. تم استخدام مجموعة بيانات **VoiceBank-DEMAND** العالمية، وهي المعيار الذهبي (Benchmark) في أبحاث فصل الكلام عن الضجيج.

1.8.3وصف مجموعات البيانات:

•**اسم الداتاست** (VoiceBank-DEMAND) **:** الإصدار المخصص لبيئة.(Kaggle

•**المصدر:** هي مزيج بين قاعدة بيانات **VCTK** (للكلام البشري النقي) وقاعدة بيانات **DEMAND** (للضجيج البيئي المتنوع)

•**المكونات :Clean Speech *:** تسجيلات صوتية لـ 28 متحدثاً بلكنات مختلفة.

• **Noisy Speech:** نفس التسجيلات السابقة ولكن مدموجة بـ 10 أنواع من الضجيج (مثل ضجيج المكتب، الكافيتريا، الشارع، وغيرها) وبمستويات مختلفة من نسبة الإشارة إلى الضجيج.(SNR)

2.8.3إحصائيات البيانات وتقسيمها :

لضمان عدم حدوث **تسريب للبيانات(Data Leakage)** ، تم اتباع استراتيجية التقسيم بناءً على **هوية المتحدث-Speaker) based Splitting)**، بحيث لا يظهر المتحدث الموجود في مجموعة التدريب ضمن مجموعة الاختبار نهائياً

آلية الفصل :تم عزل متحدثين محددين) مثل (p226, p227, p228, p230 ليتم استخدام ملفاتهم في **التحقق (Validation)** و **الاختبار (Testing)** فقط، بينما تدرب النموذج على بقية المتحدثين. هذا يضمن أن النموذج قادر على معالجة أصوات أشخاص "لم يسمعهم" أبداً أثناء التدريب.

المعيار	التفاصيل
إجمالي الملفات	آلاف المقاطع الصوتية (بصيغة .wav).
معدل العينة (Sample Rate)	16,000 هرتز (16 kHz).
توزيع التدريب	24 متحدثاً (حوالي 85% من البيانات).
توزيع التحقق/الاختبار	4 متحدثين معزولين تماماً (حوالي 15% من البيانات).

3.8.3إحصائيات البيانات وتقسيمها :

تمت معالجة الإشارات الصوتية الخام قبل إدخالها للنموذج عبر الخطوات التالية (يمكنك هنا إدراج لقطات الشاشة للكود):

1. **توحيد معدل العينة (Resampling):** تحويل كافة الملفات الصوتية إلى **16 kHz**لتقليل التكلفة الحسابية مع الحفاظ على جودة الكلام.
2. **تطبيع البيانات (Normalization):** جعل قيم الإشارة تتراوح بين [0, 1] لتسريع تقارب النموذج (Convergence)
3. **التحويل إلى الطيف الترددي(STFT):** تحويل الموجات الصوتية من النطاق الزمني إلى النطاق الترددي (Spectrograms).

9.3النموذج الأساسي:

اعتمد النموذج على معمارية **U-Net** كاملة تضم طبقاتConv2D ، BatchNorm، MaxPooling2D، وDropout ، وتم تشغيلها في بيئة Kaggle لدعم المعالجة الرسومية GPU بشكل فعال.

الخاتمة (Conclusion)

في نهاية هذا البحث، تم التوصل إلى تطوير نظام دبلجة آلي متكامل يجمع بين كفاءة تنقية الصوت ودقة الترجمة اللغوية أظهرت النتائج العملية أن استخدام بنية **U-Net** المدربة على بيانات هجينة أدى إلى تحسين جودة الصوت بشكل ملحوظ (PESQ: 2.69)، بينما ساهم دمج **WhisperX** في خفض نسبة الخطأ في الكلمات إلى مستويات قياسية (**0.02**) ، مما يضمن دقة عالية في نقل المحتوى التعليمي.

وعلى الرغم من التحديات المتعلقة بتنوع بيئات الضجيج، إلا أن النظام أثبت قدرته على العمل في الوقت الحقيقي بكفاءة زمنية عالية. نوصي مستقبلاً بزيادة حجم بيانات التدريب لتشمل لهجات عربية متنوعة، ودمج تقنيات استنساخ الصوت لرفع واقعية الدبلجة، مما يساهم بشكل فعال في كسر الحواجز اللغوية وتسهيل الوصول إلى المعرفة العالمية.

.

الفصل الرابع
التجارب والنتائج والتقييم

الفصل الرابع – التجارب والنتائج

1.4آلية عمل الاختبار:

تم اختبار النظام المقترح باستخدام مجموعة بيانات مجموعة بيانات (VoiceBank-DEMAND) بمعدل (عينة16 kHz) والتي تحتوي على تسجيلات صوتية نقية مدمجة بعشرة أنواع مختلفة من الضجيج الواقعي والبيئي، مما يسمح بتدريب النموذج على محاكاة ظروف التسجيل الحقيقية بدقة عالية.

- **المحسن (Optimizer):** تم استخدام **AdamW** لضبط الأوزان تلقائياً مع تنظيم فعال للوزن (Weight Decay)
- **معدل التعلم (Learning Rate = 0.0001):** مع تفعيل خاصية الخفض الديناميكي عند ثبات الأداء(Scheduler)
- **عدد الحقب (50 Epochs):** لنموذج U-Net لضمان الوصول لأفضل أداء دون إفراط في التدريب.
- **بيئة العمل:** تم استخدام بيئة **Kaggle** المزودة بمعالج **NVIDIA Tesla T4 GPU** لتسريع عمليات التدريب والمعالجة.

2.4نتائج الاختبار:

أظهرت النتائج أن النظام حقق أداءً استثنائياً في جميع مراحل خط المعالجة، ويمكن تلخيص النتائج الإحصائية كما يلي:

- **جودة الصوت الإدراكية (Mean PESQ):** بلغت القيمة **2.69**، مما يشير إلى تحسن كبير في جودة الصوت بعد التنقية بموديل.U-Net
- **وضوح الكلام (Mean STOI):** بلغت القيمة **0.93**، وهي نسبة عالية جداً تؤكد الحفاظ على مفهومية الكلام البشري وعدم تشوهه أثناء إزالة الضجيج.
- **جودة فصل الضجيج (Mean DNSMOS):** بلغت **3.08**، مما يعكس كفاءة النظام في التعامل مع الضوضاء المحيطة.
- **دقة التعرف على الكلام (Mean WER):** بلغت نسبة الخطأ **0.02** فقط (أي دقة تصل لـ 98%)، بفضل استخدام نموذج WhisperX

3.4تحليل الأداء الزمني:

أثبتت التجارب العملية كفاءة النظام في التعامل مع المحتوى الضخم:

- **سرعة الاستدلال:** تمكن النظام من معالجة ودبلجة فيديو تعليمي مدته **8 دقائق** في زمن قدره **12 دقائق** فقط عند استخدامGPU
- **التزامن الزمني:** بفضل المحاذاة على مستوى "الفونيم"، لم يُلاحظ أي انزياح زمني (Time Drift) حتى في الفيديوهات الطويلة، مما يضمن تطابق الصوت العربي مع حركة الفيديو الأصلي.

4.4تحليل النتائج (نقاط القوة والضعف):

أظهرت التجارب العملية توازناً كبيراً في أداء المنظومة:

1.4.4نقاط القوة:

- **دقة الترجمة والمحاذاة:** التفوق الواضح في مزامنة الصوت المترجم مع الجدول الزمني للفيديو بفضل.WhisperX
- **جودة الترجمة السياقية:** الاعتماد على النماذج اللغوية الضخمة (LLMs) منع حدوث الترجمة الحرفية الخاطئة للمصطلحات العلمية.
- **المتانة (Robustness):** قدرة نموذج U-Net على تنقية الصوت حتى في حال وجود ضجيج مروحة أو ضوضاء خلفية في المحاضرة.

2.4.4نقاط الضعف:

- **الاعتماد على العتاد:** يتطلب النظام معالجات GPU قوية ليعمل بالسرعة المطلوبة، وقد يتباطأ الأداء على الأجهزة ذات الموارد المحدودة.
- **تداخل الأصوات:** قد يواجه النظام تحديات في حال وجود أكثر من متحدث في نفس اللحظة بشكل متداخل جداً (Overlapping Speech)

5.4مقارنة النتائج:

- **مقارنة بالأنظمة التقليدية:** تفوق النظام المقترح في سرعة الدبلجة ودقة التزامن؛ حيث تعاني الأنظمة التي تعتمد على المترجمات الإحصائية من ضعف في فهم السياق التعليمي وفقدان التزامن بمرور الوقت.
- **مقارنة بنماذج Whisper القياسية:** قدم WhisperX نتائج أفضل بكثير في تحديد "الطوابع الزمنية" الدقيقة، مما ألغى الحاجة للتدخل البشري لتصحيح مواضع الصوت.

6.4التحسينات المطلوبة والتوجهات المستقبلية:

لتطوير النظام في المراحل القادمة، يمكن التركيز على:

- **استنساخ الصوت (Voice Cloning):** دمج موديلات تتيح دبلجة الفيديو بنفس نبرة صوت المحاضر الأصلي لزيادة الواقعية.
- **دعم اللهجات:** توسيع قاعدة البيانات لتشمل التعرف على اللهجات العربية المختلفة وتبسيط اللغة الأكاديمية.
- **مزامنة الشفاه (Lip-Sync):** استخدام تقنيات الـ GANs لتعديل حركة الفم في الفيديو لتتطابق تماماً مع النطق العربي الناتج.

7.4الخلاصة:

تؤكد نتائج هذا الفصل نجاح النظام في تحقيق الأهداف المرجوة حيث أثبتت مقاييس **PESQ** و **WER** أن المنظومة قادرة على إنتاج دبلجة تعليمية عالية الجودة يمثل هذا العمل أساساً قوياً لتطبيقات تعريب المنصات التعليمية المفتوحة وتقليل الفجوة اللغوية في الوصول للمعرفة العالمية.

الفصل الخامس
الخاتمة والتوصيات المستقبلية

الفصل الخامس -الخاتمة والآفاق المستقبلية

1.5الخاتمة:

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.

2.5الآفاق 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. **جمد 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.

3.5التوصيات:

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.

4.5الملاحظات 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