



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

إعداد 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، وال الحاجة لنظام دبلجة سريعة تتجاوز عقبات التكلفة والزمن في الطرق التقليدية.

تكمن بالإضافة النوعية لهذا البحث في بناء وتدريب نموذج شبكة عصبية عميقa تعتمد على بنية **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 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 دقيقة باستخدام معالج رسومي GPU ، مما يجعله حلًا فعالاً وقابلًا للتوسيع في تعريب منصات التعليم المفتوح

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

---

## **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.

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

## الفهرس

2 .....	الملخص : المحتوى
6 .....	فهرس المصطلحات
9 .....	الفصل الاول - المقدمة
9 .....	1.1 تمهيد (Introduction)
9 .....	1.2 مشكلة البحث (Problem Statement)
9 .....	1.3 أهداف البحث (Research Objectives)
9 .....	1.4 النتائج والإسهامات (Contributions and Results)
10 .....	1.5 هيكلية البحث (Research Structure)
12 .....	الفصل الثاني - الأسس التقنية للدبجة والأدبيات السابقة
12 .....	2.1 المفاهيم الأساسية (Fundamental Concepts)
12 .....	2.2 تطور الدبجة من التقليدية إلى الذكاء الاصطناعي
12 .....	2.3 علاقة الذكاء الاصطناعي بمجال الدبجة
12 .....	2.4 الأدبيات السابقة
13 .....	جدول(1-2): مقارنة تقيية بين منهجيات ومعايير قياس الأبحاث السابقة
16 .....	الفصل الثالث - التقنيات والأدوات المستخدمة
16 .....	3.1 المفاهيم التقنية والخصائص الأساسية
16 .....	3.1.1 تقنية Demucs لفصل المصادر الصوتية (Audio Source Separation)
16 .....	3.1.2 نموذج ResUNet المطور (الموديل الخاص بنا)
16 .....	3.1.3 تقنية WhisperX للمحاذة والتعرف على الكلام
16 .....	3.1.4 المقارنة بين Whisper (الأصلي) و WhisperX (المطور)
17 .....	جدول(3-1):
17 .....	مقارنة بين whisper & whisperx
17 .....	3.2 التحليل التقني لخوارزمية الشبكة العصبية التلافية (U-Net/CNN):
17 .....	3.2.1 مقدمة حول الشبكة التلافية:
17 .....	3.2.2 بنية النموذج وتقسيم الطبقات المستخدمة:
18 .....	شكل(1-3): بنية الشبكة العصبية
19 .....	3.2.3 التحسينات التي قدمها هذا المنهج مقارنة بالأساليب التقليدية:
21 .....	الفصل الرابع: التجارب والنتائج والتقييم
21 .....	4.1 التحديات والاحتياجات (Challenges and Requirements)
21 .....	4.2 مجموعة المعطيات المستخدمة وتحضيرها (Datasets)

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

---

21.....	جدول (4-1): إحصائيات تقسيم مجموعة البيانات المستخدمة .....
22.....	4.3 مخطط المنهجية العام (Methodology Framework): .....
22.....	الشكل (4-1): المخطط الانسيابي العام لمنهجية عمل نظام الدبلجة الآلي .....
23.....	4.4 آلية عمل الاختبار والتدريب: .....
23.....	شكل (4-2): منحنيات تطور دالة الخسارة ومقاييس SI-SDR أثناء تدريب النموذج .....
23.....	4.5 مقاييس التقييم: .....
23.....	جدول (4-2): مقاييس تقييم الأداء المستخدمة في المشروع .....
24.....	4.6 نتائج الاختبار وتحليل الأداء: .....
24.....	الشكل (4-3): النتائج الإحصائية النهائية لأداء نظام الدبلجة الآلي .....
26.....	الفصل الخامس - الخاتمة والأفاق المستقبلية .....
26.....	5.1 الخاتمة: .....
26.....	5.2 الأفاق المستقبلية: .....
27.....	5.3 التوصيات: .....
27.....	5.4 الملاحظات الخاتمية: .....
28.....	المراجع References

## فهرس المصطلحات

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

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

المعنى	الاختصار	الترجمة العربية	المصطلح التقني (بالإنجليزية)
المقياس الأساسي لتقدير دقة أنظمة التعرف على الكلام (ASR)، حيث يقيس الفرق بين النص الناتج والنص الأصلي 1	<b>WER</b>	نسبة الخطأ في الكلمة	<b>Word Error Rate</b>
دالة خسارة تستخدم غالباً في تدريب شبكات-U Net لقياس الفرق بين الصوت النقي والصوت الناتج عن النموذج.	<b>MSE</b>	متوسط مربع الخطأ	<b>Mean Squared Error</b>
دالة رياضية تقيس مدى دقة تنبؤات النموذج أثناء عملية التدريب لتقليل الخطأ تدريجياً.	-	دالة الخسارة	<b>Loss Function</b>
خوارزمية (مثل Adam) تُستخدم لتحديث أوزان الشبكة العصبية لتقليل دالة الخسارة أثناء التدريب	-	المحسن	<b>Optimizer</b>
الوقت الذي يستغرقه النموذج لمعالجة البيانات (مثل الفيديو) وإخراج النتيجة النهائية.	-	زمن الاستدلال	<b>Inference Time</b>
بنية هندسية تعتمد عليها شبكة U-Net؛ حيث يقوم المشفر بضغط البيانات وفك التشفير بإعادة بنائها	-	المشفر وفك التشفير	<b>Encoder-Decoder</b>
ميزة في بنية U-Net تسمح بنقل المعلومات مباشرةً بين الطبقات المقابلة لحفظ التفاصيل الدقيقة للصوت	-	وصلات التخطي	<b>Skip Connections</b>
جزء من البيانات يستخدم لاختبار النموذج أثناء التدريب للتأكد من قدرته على التعميم ومنع فرط التخصيص	-	مجموعة التحقق	<b>Validation Set</b>

# **الفصل الأول**

## **المقدمة**

## **الفصل الأول - المقدمة**

### **1.1 تمهيد (Introduction)**

يشهد العصر الحالي تحولاً رقمياً هائلاً في مجال التعليم المفتوح، حيث أصبحت المنصات التعليمية العالمية مصدرًا 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.

### **1.2 مشكلة البحث (Problem Statement)**

تواجده أنظمة دبلجة الفيديوهات التقليدية والأالية عدة تحديات تعيق اعتمادها بشكل واسع، أبرزها تأثير جودة التعرف على الكلام بوجود الضجيج والتشویش في التسجيلات الأصلية، بالإضافة إلى صعوبة مطابقة النص المترجم مع الطوابع الزمنية للفيديو الأصلي بشكل دقيق (دقة التزامن). كما تبرز مشكلة التكلفة والזמן، حيث تتطلب الدبلجة البشرية موارد مالية ضخمة ووقتاً طويلاً للمعالجة. يعالج هذا البحث هذه المشكلات عبر بناء منظومة متكاملة تبدأ بموデل U-Net مدرب خصيصاً لفصل الضجيج، متبعاً بنظام WhisperX لضمان الترجمة الدقيقة والمحاذاة الزمنية الاحترافية.

### **1.3 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 لتناسب حركة الشفاه والفيديو الأصلي باستخدام تقنيات المحذاة على مستوى "الفنون".
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)

### **1.4 النتائج والإسهامات (Contributions and Results)**

قدم البحث حلولاً تقنية مبتكرة في مجال المعالجة الصوتية واللغوية، حيث تم بناء وتدريب نموذج U-Net عصبي 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 "الفنون" (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 دبلجة فيديو تعليمي مدمج GPU ، مما يعزز من إمكانيات استخدامه في تطبيقات ترجمة المحتوى التعليمي الضخم بشكل آلي وسريع.

### **1.5 هيكلية البحث (Research Structure)**

يتألف هذا البحث من ستة فصول متراابطة 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 بالقصص. 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 تأثير المفاهيم الأساسية التي يقوم عليها هذا البحث:

- **الإشارة الصوتية (Audio Signal):** تمثل فيزيائيًّا للأصوات في النطاق الزمني أو التردد، 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 ثلثينيات 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 مكافحة 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 حلوًّا 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 عصبية مثل U-Net لتنقية الصوت من الضجيج.
2. **المحاداة الزمنية الدقيقة :** باستخدام تقنيات المحاداة 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.

### **2.4 الأدبيات السابقة**

تنوعت 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 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 المحاداة القسرية لل Phoneme.

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

- لضمان توقيتات دقة على مستوى الكلمة، مما عالج مشكلات "الهلوسة" (Forced Phoneme Alignment) والانزياح الزمني محققة تسارعاً يصل إلى 12 ضعفاً [1].
- دراسة "Pyannote.audio: neural building blocks for speaker diarization" [2]** استعرضت نظام `pyannote.audio` كمنهجية متقدمة لتقسيم وتحديد المتحدثين في البيانات المعددة؛ إلا أن مشروعنا ركز على سيناريو المتحدث الواحد (Single Speaker)، مما سمح بتبسيط هذه المرحلة وتوجيهه موارد المعالجة لضمان استقرار الهوية الصوتية للمتحدث الوحيد [2].
  - دراسة "Language-guided audio-visual source separation via trimodal consistency" [3]** 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" [4]** ركزت على XTTS لاستنساخ البصمة الصوتية (Voice Cloning) بدقة عالية، ومع ذلك، اعتمد مشروعنا على تقنية Edge TTS التي تفوقها في جودة اللغة العربية وسرعة الاستجابة (Low Latency) وتوفير نبرات طبيعية دون الحاجة لمتطلبات عناية ضخمة [4].
  - دراسة "A lip sync expert is all you need for speech to lip generation in the wild" [5]** قدمت نموذج Wav2Lip الذي يعتمد على "ميفر خبير" (Expert Discriminator) لاكتشاف عدم التطابق بين الصوت وحركة الشفاه، ويمثل هذا النموذج المعيار التقني المستهدف في مشروعنا لضمان تطابق حركة الشفاه مع النص العربي المولد في مراحل التطوير المتقدمة [5].

ويمكن تلخيص الدراسات المرجعية في الجدول التالي :

جدول(2-1): مقارنة تقنية بين منهجيات ومعايير قياس الأبحاث السابقة

معايير التقييم (Metrics)	آلية العمل (Methodology)	العام	التقنية (Technology)	المراجع
نسبة WER (نسبة خطأ الكلمات)، عامل RTF (الوقت الحقيقي)، دقة الطوابع الزمنية.	يعتمد استراتيجية Phoneme Alignment عبر دمج نموذج Whisper مع نماذج فونيمية (مثل wav2vec2) لربط النص بالصوت بدقة زمنية عالية.	2023	WhisperX	[1]
معدل DER (معدل خطأ التقسيم)، و معدل JER (معدل جاكارد) لتقدير دقة تحديد هوية المتحدث.	بنية عصبية تعتمد على EEND-VC للمتحدثين، وتستخدم تقنيات العنقودة (Clustering) لتحديد هوية كل متحدث في المقطع.	2020	pyannote.audio	[2]

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

معايير التقييم (Metrics)	آلية العمل (Methodology)	العام	التقنية (Technology)	المراجع
نسبة SDR (الإشارة إلى التشوه)، و CLAPScore لتقدير الصلة الدلالية بين النص والصوت المعزول.	يعتمد على Trimodal Consistency (الاتساق الثلاثي) بين الصوت، اللغة، والرؤية عبر تقنية التعلم التباعي (Contrastive Learning) لعزل المصادر.	2023	VAST	[3]
نسبة CER (خطأ الحروف)، تشابه SECS البصمة الصوتية)، و UTMOS للتقدير الذاتي للجودة.	نموذج لغوي (GPT-) based Conditioning Encoder لاستخلاص بصمة المتحدث Zero-"shot" وتوليد كلام بأسلوب "shot" عبر لغات متعددة.	2024	XTTS	[4]
LSE-D (مسافة خط المزامنة)، LSE-C (ثقة المزامنة)، و FID لتقدير جودة الصورة.	يستخدم شبكة GAN مع مميز Expert (خبير) Discriminator (مدرس) مسبقاً SyncNet (لفرض دقة حركة الشفاه وتوافقها مع الكلام المولد).	2020	Wav2Lip	[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 (Audio Source Separation)**

• ما هو Demucs؟ هو نموذج ذكاء اصطناعي متتطور 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 بشرى نقى 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 عكس Demucs ، 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".

##### **3.1.3 تقنية WhisperX للمحاداة والتعرف على الكلام:**

• ما هو WhisperX؟ هو نسخة "محسنة" ومطورة من نموذج Whisper الأصلي. هو ليس مجرد أداة لتحويل الصوت إلى نص، بل هو نظام متكامل للمحاداة 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 المدخل 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(3-1)**

#### **مقارنة بين whisper & whisperx**

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

### **3.2 التحليل التقني لخوارزمية الشبكة العصبية التلaffيفية (U-Net/CNN)**

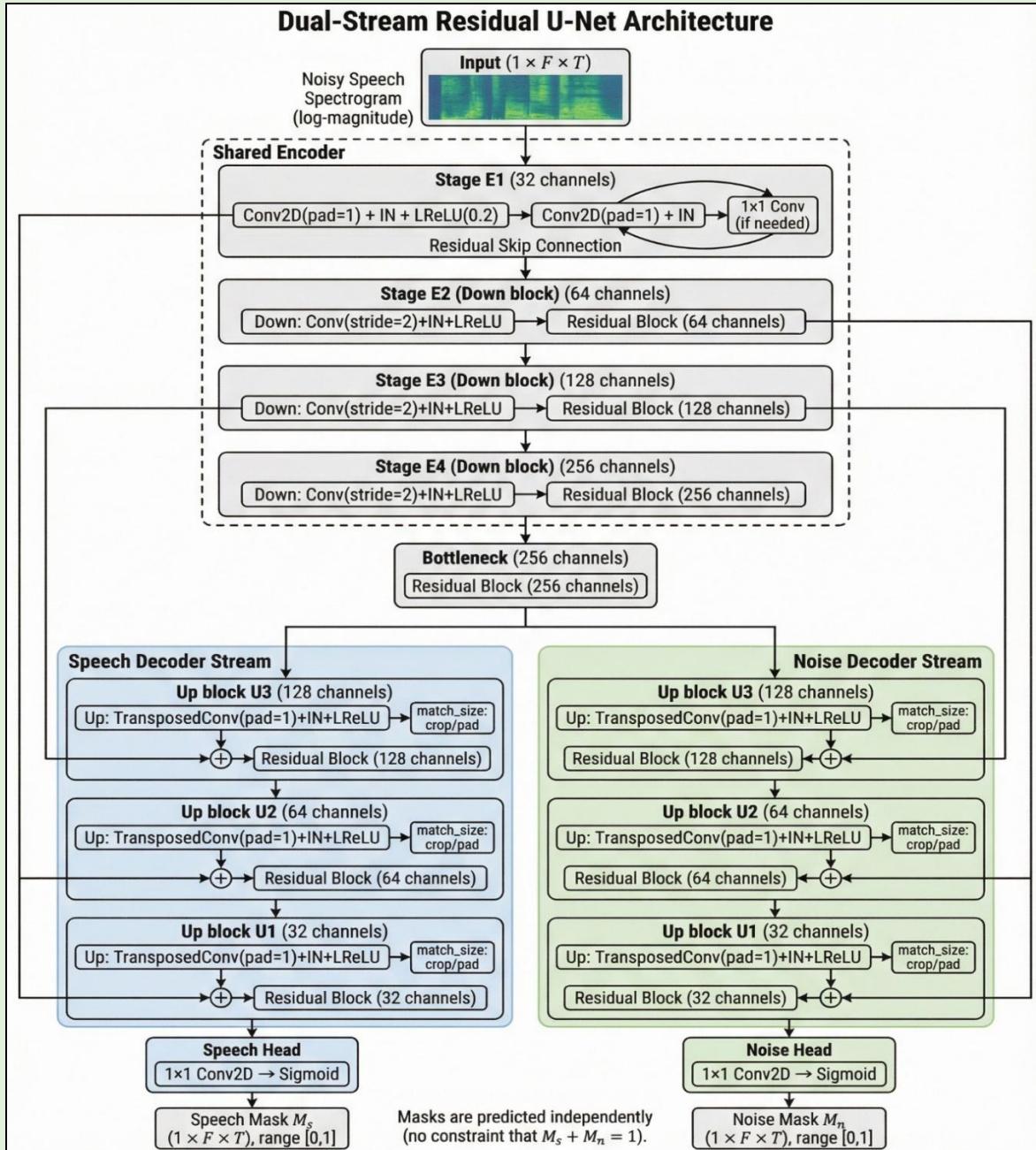
#### **3.2.1 مقدمة حول الشبكة التلaffيفية:**

تعتبر الشبكات التلaffيفية CNN - U-Net (Convolutional Neural Networks) وبنيتها المتطرورة الأنسب لمعالجة الإشارات الصوتية بعد تحويلها لمخططات طيفية، حيث تعتمد فكرة العمل على استخدام طبقات تلaffيفية (Convolutional Layers) لاستخراج الميزات الهرمية من الترددات الصوتية، مما يجعلها فعالة للغاية 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) الهيكالية التفصيلية لهذه البنية، حيث نلاحظ كيف يقوم المشفر باستخدام الميزات بينما تعمل طبقات فك التشفير على إعادة بناء الإشارة الصوتية النقية مع الحفاظ على التفاصيل عبر وصلات التخطي.

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



شكل (3-1): بنية الشبكة العصبية

وت تكون هذه البنية من:

- **الطبقات التلافيافية المتبقية (Residual Convolutional Layers):**
  - تستخدم لاستخراج الميزات العميقة من بيانات المخطط الطيفي عبر مرشحات متضاعفة (32, 64, 128).
  - تعتمد على دالة التنشيط **Leaky ReLU** لضمان تدفق التدرجات ومنع مشكلة "تلاثي 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 ويبعد فقدان المعلومات أثناء الضغط.
- **طبقات التقسيس والانتظام (Batch Normalization & Dropout):**
  - تستخدم **Batch Normalization** لتحقيق استقرار التدريب وسرعة التقارب.
  - تستخدم طبقات **Dropout** بنسبة محددة لمنع التكيف الزائد (Overfitting)، مما يجعل النموذج قادرًا على التعامل مع أنواع ضجيج لم يراها مسبقاً.
- **طبقة إعادة البناء (Decoder & Upsampling):**
  - تقوم بإعادة بناء الإشارة الصوتية النقية، وتستخدم في نهايتها طبقة تلافيافية لتوليد القاتع (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. **تقدير أخطاء التزامن (Reduced Misclassifications):** دمج **WhisperX** يضمن مطابقة النص المترجم مع الطابع الزمني للفيديو بدقة "الفونيم"، مما يلغي التداخل الزمني.
3. **سرعة الأداء (Optimized Speed):** المعالجة المباشرة للمخططات الطيفية تتيح دبلجة فيديو مدته 8 دقائق في غضون 12 دقيقة 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**

#### **الفصل الرابع: التجارب والنتائج والتقييم**

يسعرض هذا الفصل الجوانب التطبيقية والمنهجية لنظام الدبلجة الآلي المقترن، 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 بيئه الاختبار.

#### **4.1 التحديات والاحتياجات:(Challenges and Requirements)**

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

- **مجموعات المعطيات:** تطلب الشبكات التلافيية بيانات صخمة ومتعددة للتكييف 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 مجموعة المعطيات المستخدمة وتحضيرها :(Datasets)**

لضمان كفاءة النموذج في الظروف الواقعية، تم الاعتماد على مجموعة بيانات **VoiceBank-DEMAND** المعيارية. تتميز هذه المجموعة بتغطية واسعة لأنواع الضجيج المختلفة المسجلة في بيانات حقيقة (مثل ضجيج المكاتب، المطبخ، والشارع) مدمجة مع كلام بشري نقى من قاعدة بيانات VCTK يوضح الجدول (4-2) التوصيف الإحصائي الموسع للبيانات المستخدمة في التدريب والاختبار:

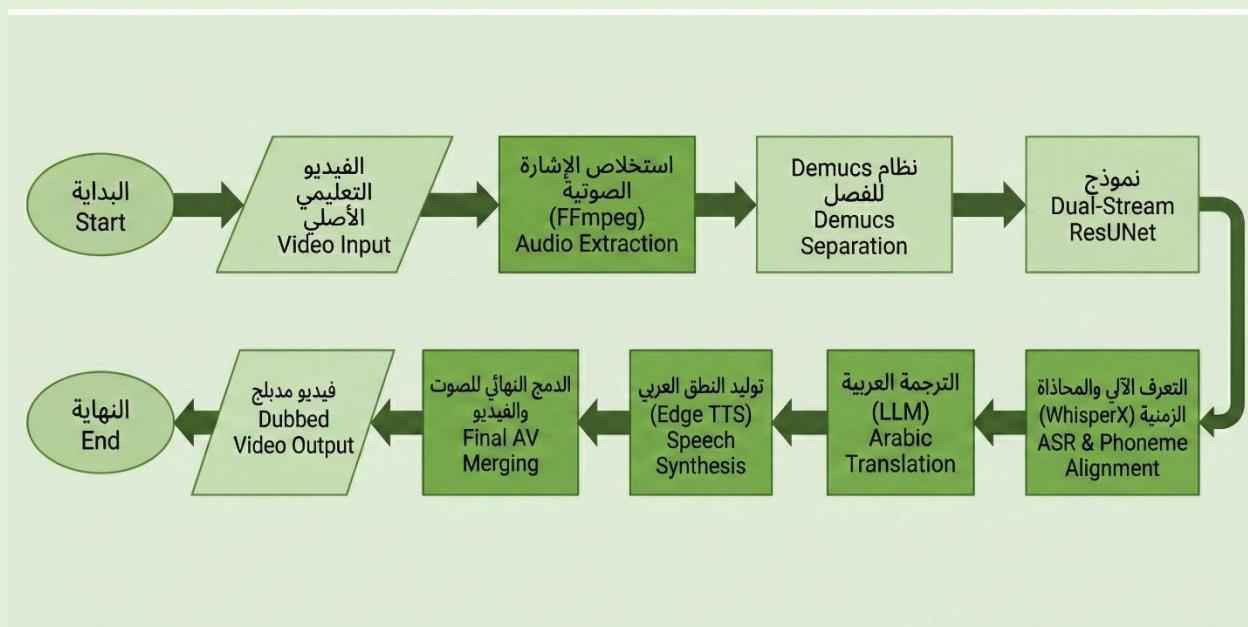
**جدول (4-1): إحصائيات تقسيم مجموعة البيانات المستخدمة**

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 معزولاً تماماً لضمان حيادية التقييم.
معدل العينة (Sampling Rate)	موحد عند 16,000 هرتز لضمان استقرار المعالجة التلافيية.

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

### 4.3 مخطط المنهجية العام : (Methodology Framework)

يوضح الشكل (4-1) المسار التسلسلي لمنهجية عمل النظام، حيث تتبع عملية المعالجة خطوات تقنية محددة لضمان جودة الدبلجة النهائية.



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

وتلخص وظائف المكونات المنهجية فيما يلي:

- استخلاص الإشارة الصوتية (Audio Extraction): باستخدام أداة FFmpeg ، يتم فصل المسار الصوتي عن الفيديو الأصلي بصيغة WAV وبمعدل عينة 16 kHz لضمان توافق البيانات مع النماذج اللاحقة.
- نظام الفصل (Demucs Separation): تعمل هذه المرحلة على عزل الموسيقى الخلفية والآلات الموسيقية عن المسار الصوتي، مما يترك كلاماً بشرياً أو لياً يحتاج لمزيد من التقطة.
- نموذج (Dual-Stream ResUNet): وهو المكون الجوهرى المطور في البحث، حيث يقوم بتنقية الكلام من الضجيج البيئي والبشري والصدى، لإنتاج إشارة صوتية نقية تماماً جاهزة لمرحلة التعرف.
- التعرف الآلي والمحاذاة الزمنية (WhisperX): تحويل الصوت النقي إلى نص مكتوب (ASR) مع استخراج طوابع زمنية دقيقة على مستوى "الفنون" لضمان تزامن مثالي لاحقاً.
- الترجمة العربية (Arabic Translation): استخدام نماذج لغوية ضخمة (LLM) لترجمة النص المستخلص من الإنجليزية إلى العربية مع الحفاظ على الدقة العلمية والسياق التعليمي.
- توليد النطق العربي (Edge TTS): تحويل النص المترجم إلى صوت بشري طبيعي بجودة عالية، معضبط سرعة النطق لتتوافق مع المدد الزمنية الأصلية.
- الدمج النهائي (Final AV Merging): إعادة تركيب المسار الصوتي العربي الجديد مع الفيديو الأصلي وتصدير المنتج النهائي بجودة عالية.

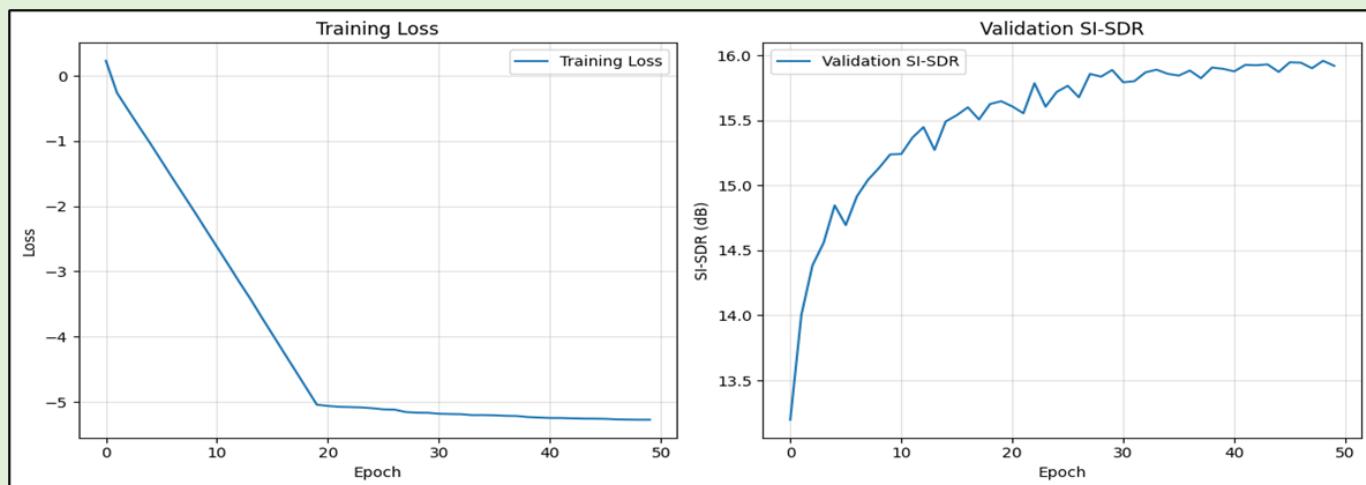
## Automated Dubbing System for Educational Videos using Deep Learning Techniques (English to Arabic)

### 4.4 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 المسماقة:** توحيد معدل العينة إلى 16 kHz وتحويل الموجات إلى نطاق التردد (STFT).

ويوضح الشكل (4-2) منحنيات الأداء أثناء عملية التدريب، حيث نلاحظ الانخفاض التدريجي والمستقر في دالة الخسارة (Training Loss) ، بالتزامن 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 (4-2): منحنيات تطور دالة الخسارة ومقياس SI-SDR أثناء تدريب النموذج

### 4.5 مقاييس التقييم:

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

جدول (4-2): مقاييس تقييم الأداء المستخدمة في المشروع

الوصف والهدف	المقياس
لتقييم جودة الصوت الإدراكية ومدى قربه من الصوت النقي الأصلي.	PESQ

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

الوصف والهدف	المقياس
لقياس مدى وضوح وفهم الكلام البشري بعد المعالجة.	STOI
لقياس دقة التعرف على الكلام عبر حساب نسبة الكلمات الخاطئة.	WER
مقياس موضوعي لجودة فصل الضجيج في البيئات الواقعية.	DNSMOS

### 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 (4-3) 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 (4-3): النتائج الإحصائية النهائية لأداء نظام الدبلجة الآلي

```
=====
🏆 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
```

## **الفصل الخامس**

# **الخاتمة والتوصيات المستقبلية**

## **الفصل الخامس - الخاتمة والآفاق المستقبلية**

### **5.1 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 المعالجة وفصل الضجيج الصوتي، وتقنيات WhisperX للمحاداة 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 موضوعات خفية.

بلغت دقة النموذج في التعرف على الكلام (WER) نسبة خطأ ضئيلة جداً ووصلت 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 مثلمحاكاة نبرة الصوت الأصلية (Voice Cloning) والتعامل مع التداخلات الصوتية المعقّدة في الفيديوهات ذات الجودة المتدنية جداً.

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

### **5.2 الآفاق المستقبلية:**

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

#### **1. تحسين جودة استنساخ الصوت (Voice Cloning):**

- تطوير تقنيات متقدمة تتبع 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 المعقّدة بدقة أعلى.

#### **3. استخدام تقنيات تعلم عميق هجينه:**

- تجربة دمج بنى شبكات عصبية أكثر تطوراً مثل **Transformers** لمعالجة الصوت أو شبكات **GANs** لتحقيق مزامنة بصيرية (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.

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

---

### **5.3 التوصيات:**

1. **تحسين جودة المعطيات:** لتعزيز دقة النموذج في مختلف الظروف، يجب التركيز على تنوع بيانات التدريب بما يعكس تعقيدات العالم الحقيقي (مثل صدى القاعات، ضجيج الأجهزة، ونقاوت جودة الميكروفونات).
2. **تطوير أدوات تحليلية مرنية:** 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 وفهمًا للسياق الثقافي العربي.

### **5.4 الملاحظات الختامية:**

يشكل هذا البحث خطوة مهمة نحو استخدام الذكاء الاصطناعي في كسر الحاجز اللغوي للوصول إلى المعرفة العالمية ومع تقدم تقنيات التعلم العميق، يمكن تحسين دقة المنظومة وجعلها أكثر كفاءة في البيانات 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**