مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (٩) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (9) 2025

تحسين أنظمة التعرف التلقائي على الكلام باستخدام نموذج Conformer وتقنيات تقليل السمات

د. م. راغب طعمه*

م. احمد حكمات محمد**

(تاريخ الإيداع ٢٠٢٥/٦/١٦ . قُبِل للنشر في ٢٠٢٥/٨/٤)

□ ملخّص□

يُعد التعرف التلقائي على الكلام (Automatic Speech Recognition – ASR) من المجالات الرائدة في تقنيات الذكاء الاصطناعي والتعلم العميق، لما له من تطبيقات واسعة في المساعدات الرقمية، وتحويل الكلام إلى نص، والتفاعل الصوتي في الأجهزة الذكية. تهدف هذه الدراسة إلى تقديم إطار منهجي محسن لتعزيز كفاءة ودقة أنظمةASR ، بالاعتماد على تقنيات معالجة صوتية متقدمة وبنية نمذجة هجينة فعالة.

يعتمد النظام المقترح على استخدام طيف الطاقة لبنوك المرشحات (Filter Bank Energies – Fbank) كبديل عن المعاملات الطيفية التقليدية مثل(Mel–Frequency Cepstral Coefficients – MFCC) ، لما توفره من معلومات طيفية دقيقة تساعد على تحسين تمييز الأنماط الصوتية. كما تم توظيف تقنيةSpecAugment ، القائمة على التحوير الزمني والترددي، بهدف زيادة تنوع البيانات المُستخدمة في التدريب وتعزيز قدرة النموذج على التعميم في بيئات صوتية متنوعة. في بناء النموذج، تم اعتماد بنيةConformer ، وهي بنية هجينة تدمج بين الشبكات الالتفافية Convolutional) Neural Networks – CNNs)والمحوّلات(Transformers) ، مما يُمكّن النموذج من التقاط الأنماط الصوتية الزمنية والمحلية والعالمية بكفاءة أعلى. وقد تميز النظام المقترح كذلك بتقليل عدد السمات الصوتية إلى ٥٣ سمة فقط، مما أسهم في تقليل التعقيد الحسابي وتقليل استهلاك الموارد، دون التأثير سلبًا على الأداء.

أظهرت النتائج التجريبية تفوق النموذج من حيث الكفاءة والدقة، حيث بلغ معدل الخطأ في الكلمات (WER) نحو ١٩%، مع وصول قيمة الخسارة (Validation Loss) إلى ٠,٢١. وتؤكد هذه النتائج أن النظام المقترح قادر على التعامل مع تحديات بيانات الصوت الواقعية، ويُمثّل خطوة واعدة نحو تحسين أداء أنظمة التعرف التلقائي على الكلام. كما تُمهّد هذه الدراسة الطريق لمزيد من الأبحاث المستقبلية التي تستهدف تحسين البنى المعمارية للنماذج ودمج تقنيات تعلم جديدة.

الكلمات المفتاحية :التعلم العميق، التعرف التلقائي على الكلام، الشبكات العصبونية، معدل الخطأ في الكلمات، تخفيض السمات، Fbank ، Conformer.

*أستاذ مساعد في قسم هندسة تكنولوجيا المعلومات – كلية هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس – سوريا.

** حاصل على الماجستير في هندسة تكنولوجيا المعلومات–كلية هندسة تكنولوجيا المعلومات والاتصالات–جامعة طرطوس–سوريا.

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (٩) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (9) 2025

Improving Of Automatic Speech Recognition Systems Using By Conformer Model And Feature Reduction Techniques

***Dr. Ragheb Toghmah**
****Ahmad Hikmat Mohammad**

(Received 16/6/2025 . Accepted 4/8/2025)

□ ABSTRACT □

Automatic Speech Recognition (ASR) is a leading field in artificial intelligence and deep learning technologies, with broad applications in digital assistants, speech-to-text conversion, and voice interaction in smart devices. This study aims to present an improved methodological framework to enhance the efficiency and accuracy of ASR systems, relying on advanced audio processing techniques and an effective hybrid modeling architecture.

The proposed system relies on the use of the energy spectrum of filter banks (Fbanks) as an alternative to traditional spectral coefficients such as MFCC, as it provides accurate spectral information that helps improve the recognition of audio patterns. The SpecAugment technique, based on temporal and frequency convolution, was also employed to increase the diversity of data used in training and enhance the model's ability to generalize across diverse audio environments. To build the model, the Conformer architecture was adopted, a hybrid architecture that combines Convolutional Neural Networks (CNNs) and Transformers, enabling the model to capture temporal, local, and global acoustic patterns more efficiently. The proposed system also features a reduction in the number of acoustic features to only 53, which contributes to reducing computational complexity and resource consumption without negatively impacting performance.

Experimental results demonstrated the model's superiority in terms of efficiency and accuracy, with a Word Error Rate (WER) of approximately 19%, with a Validation Loss of 0.21. These results confirm that the proposed system is capable of handling the challenges of real-world audio data and represent a promising step toward improving the performance of automatic speech recognition systems. This study also paves the way for future research aimed at improving model architectures and incorporating new learning techniques.

Key word:Deep learning ,Automatic speech recognition (ASR),Neural Networks,word error Rate (WER),Feature Reduction,Fbank ,Conformer

* Associate Professor, Department of Information Technology Engineering, Information and Communication Technology Engineering, Tartous University, Syria.
** Master, Department of Information Technology Engineering, Information and Communication Technology Engineering, Tartous University, Syria.

١ – مقدمة:

يُعتبر التعرف التلقائي على الكلام (ASR) من المجالات الحيوية في البحث العلمي الحديث، حيث يهدف إلى تطوير تقنيات متقدمة لتحويل اللغة المنطوقة إلى نص مكتوب بدقة وفعالية. شهد هذا المجال تطورات كبيرة خلال السنوات الأخيرة، ساعدت في انتشار تطبيقات متعددة تشمل المساعدين الافتراضيين، وأنظمة التحكم الصوتي، وخدمات النسخ الفوري، وترجمة اللغات. ويكمن جوهر التعرف التلقائي على الكلام في تصميم خوارزميات ونماذج ذكية قادرة على التعامل مع تعقيدات إشارة الكلام وتحويلها إلى تمثيلات نصية دقيقة، مما يعزز التواصل بين الإنسان والآلة.

يتميز هذا المجال بطبيعته متعددة التخصصات، حيث يستفيد من علوم اللغويات، ومعالجة الإشارات الصوتية، والتعلم الآلي، وعلوم الحاسوب. كما يواجه الباحثون تحديات عدة، مثل اختلاف لهجات المتحدثين، التنوع اللغوي، الضوضاء الخلفية، وسرعة الكلام، بالإضافة إلى ضرورة تحقيق أداء فعال في بيئات الوقت الحقيقي [1].

في السنوات الأخيرة، برزت تقنيات التعلم العميق كأداة فعالة لتحسين أداء أنظمة التعرف على الكلام، لما توفره من قدرة على تعلم التمثيلات الهرمية المعقدة لإشارة الكلام، بدءاً من الميزات الصوتية الأساسية مثل الطيف الترددي، إلى تمثيلات عالية المستوى تعكس الصوتيات والكلمات. في هذا السياق، تبرز أهمية استخدام ميزات طيف الطاقة لبنوك المرشحات (Fbank)كبديل عن معاملات MFCC التقليدية [2] ، لما تقدمه من معلومات طيفية أكثر شمولاً. كما تلعب تقنيات زيادة البيانات مثل SpecAugment دوراً مهماً في تعزيز قدرة النماذج على التعميم من خلال تحوير البيانات التدريبية زمنياً وترددياً، مما يقلل من الإفراط في التخصيص.

علاوة على ذلك، يُعد نموذجConformer ، الذي يجمع بين مزايا الشبكات الالتفافية (Convolution) وتقنية المحولات (Transformer)، أحد أحدث البنى المعمارية التي تُحسن من استيعاب الأنماط الصوتية طويلة وقصيرة المدى بشكل متكامل، مما يُسهم في رفع دقة وكفاءة أنظمة التعرف التلقائي على الكلام.

تهدف هذه الدراسة إلى استكشاف وتحليل تأثير استخدام Fbank، وتقنية SpecAugment، وبنية Conformer في تطوير نظام تعرّف تلقائي على الكلام قادر على تقديم أداء متقدم وموثوق في تطبيقات العالم الحقيقي.

٢ – الدراسات السابقة:

في عام ٢٠٠١، قام الباحثون Nadeu, C. et al. [٣] بتطوير نظام تعرف تلقائي على الكلام يستخدم ميزات طيف الطاقة لبنوك المرشحات (Fbank) مع بنية نموذج تعتمد على الشبكات الالتفافية والمحولات (Conformer). أظهرت الدراسة تفوق هذا النموذج في التعامل مع الضوضاء الخلفية مقارنةً بالنماذج التقليدية المعتمدة على MFCC، وحقق معدل خطأ بالكلمة (WER) أقل على مجموعة بيانات. وقد أبرزت نتائج هذه الدراسة أهمية اعتماد تمثيلات طيفية أغنى مثل Fbank لتحسين أداء الأنظمة الصوتية في بيئات غير مثالية.

في الدراسة التي أجراها Jin, Z. et al. عام ٢٠٢٤ [٤]، تم استخدام تقنية زيادة البيانات SpecAugment لتحسين أداء نموذج ASR مبني على بنية Conformer. أظهرت النتائج تحسنًا ملحوظًا في دقة التعرف على الكلام، خصوصًا في بيئات الضوضاء والصوتيات المتغيرة، حيث ساهمت تقنية SpecAugment في تقليل معدل الخطأ بنسبة تزيد على ٢٠% مقارنةً بالنماذج التي لم تستخدم هذه التقنية. وأشارت الدراسة أيضًا إلى أن إدخال تنوع زمني وترددي أثناء التدريب يُعزز من قدرة النموذج على التعميم ومعالجة لهجات متعددة.

أما في عام ٢٠٢٢، فقد قام الباحث Liu, M. et al. [٥] بتقييم أداء نموذج Conformer في التعرف على الكلام مع مقارنة بين ميزات MFCC وFbank. أظهرت النتائج تفوق استخدام Fbank في تحسين جودة التمثيلات الصوتية، مما أدى إلى تحسين عام في دقة النموذج، وتحقيق معدل خطأ بالكلمة (WER) أقل على بيانات TED-LIUM 3. كما أوضحت الدراسة أن Fbank يوفر معلومات طيفية أكثر استقرارًا وأقل تأثرًا بالضجيج، مما يجعله خيارًا فعالًا في تطبيقات ASR الواقعية.

وفي دراسة أخرى لعام ٢٠٢٢، استخدم Nghia, H. et al. [٦] نموذج Conformer مدمجًا مع SpecAugment وميزات Fbank لتطوير نظام ASR موجه للغات ذات موارد بيانات محدودة. ساعد هذا الدمج على تحسين قدرة النموذج في التعميم والتعرف على الكلام بدقة عالية رغم قلة البيانات، حيث وصل معدل دقة التعرف إلى ٨٧% في ظروف صعبة. وأبرزت الدراسة دور البنية النموذجية المتقدمة وتقنيات المعالجة المسبقة في تعويض النقص في البيانات التدريبية.

كما أشار الباحث Geng, J. et al. [٧] في دراسته عام ٢٠٢٤ إلى أهمية الجمع بين الشبكات الالتفافية وتقنية المحولات في نموذج Conformer لتعزيز استخراج الميزات الزمانية والمكانية من إشارات الكلام، مما يؤدي إلى تحسين سرعة التعرف ودقته في الوقت الحقيقي، مع تقليل معدل الخطأ. وخلصت الدراسة إلى أن البنية الهجينة لنموذج Conformer تُعد من أكثر التصاميم كفاءة في التطبيقات التفاعلية والمعتمدة على الاستجابة اللحظية.

٣- مشكلة البحث:

يُعدّ التعرف التلقائي على الكلام تحدياً تقنياً ومعرفياً في العديد من اللغات، لا سيما تلك التي تتميز بتنوع لهجاتها وغنى مخزونها الصوتي، مما يزيد من صعوبة النمذجة الدقيقة للنطق الصوتي. تواجه أنظمة التعرف التلقائي على الكلام مشكلات ملحوظة عندما تكون البيانات الصوتية المتوفرة محدودة أو عندما تكون جودة التسجيلات متدنية أو تحتوي على ضوضاء خلفية، وهو ما يؤثر بشكل مباشر على دقة النموذج.

ومع استخدام طرق تقليدية في استخراج الميزات مثلMFCC ، تظل بعض الخصائص الصوتية غير مستغلة بالكامل، مما يحدّ من قدرة النماذج على التمييز بين الكلمات في بيئات واقعية. كما أن الأساليب التقليدية لزيادة البيانات قد تكون غير كافية لمحاكاة تنوع الظروف الحقيقية للكلام.

تظهر الحاجة إلى تطوير أنظمة أكثر مرونة ودقة من خلال استخدام ميزات صوتية أكثر تعبيراً مثل Filter Bank Energies، وتقنيات حديثة لزيادة البيانات مثل SpecAugment [8]، بالإضافة إلى اعتماد نماذج هجينة تجمع بين قدرات الشبكات الالتفافية والتحويلية كما هو الحال في Conformer. ومن هنا تنبع مشكلة البحث، حيث يُطرح التساؤل حول مدى فاعلية هذه الأساليب الحديثة في تحسين أداء أنظمة التعرف التلقائي على الكلام، خاصة في البيئات الصعبة أو ذات الموارد المحدودة.

٤- أهمية البحث وأهدافه:

يُقدم هذا البحث مساهمة نوعية في تطوير أنظمة التعرف التلقائي على الكلام عبر دمج تقنيات متقدمة في مجال معالجة الإشارات الصوتية والتعلم العميق، حيث يعتمد على استخدام مصفوفة طاقة المرشحات (Filter Bank Energies) لاستخلاص ميزات أكثر واقعية ودقة من الإشارات الصوتية، بالإضافة إلى تطبيق تقنية SpecAugment كوسيلة فعّالة لزيادة حجم وتنوع البيانات التدريبية وتحسين تعميم النموذج. كما يتميز النموذج المقترح باعتماده على بنية Conformer

التي تجمع بين إمكانيات الشبكات التلافيفية (CNN) والشبكات التحويلية (Transformer) مما يعزز من قدرته على التعرف على الأنماط الصوتية في السياقات الزمنية الطويلة والمعقدة.

تكمن أهمية هذا البحث في تحسين دقة أنظمة التعرف على الكلام خصوصاً في البيئات الواقعية التي تتسم بالتشويش أو اللهجات المتعددة، وذلك من خلال تصميم نموذج متكامل قادر على معالجة التحديات التقليدية التي تواجه هذا المجال، كالضوضاء والتنوع اللهجي وقلة البيانات.

ويهدف هذا البحث إلى تصميم نظام متقدم للتعرف على الكلام يعتمد على بنية Conformer لتحقيق أعلى مستوى من الدقة والكفاءة واستكشاف فعالية استخدام Filter Bank Energies بدلاً من الأساليب التقليدية كـ MFCC في تحسين جودة تمثيل الصوت. وبذلك يسهم هذا البحث في تطوير حلول ذكية وعملية تدعم التكامل في تطبيقات الذكاء الاصطناعي المرتبطة بالصوت، مثل المساعدات الذكية، والنظم التفاعلية، والرعاية الصحية.

٥- طرق البحث ومواده:

يعتمد هذا البحث على نموذج حديث يجمع بين البنية المعمارية المتقدمة Conformer التي تمزج بين الشبكات التلافيفية (CNN)وآليات الانتباه الذاتي (Self-Attention) المستمدة من التحويلات(Transformers) [9]، وذلك بهدف تحقيق توازن بين التقاط الأنماط الزمنية المحلية والعلاقات السياقية طويلة المدى في الإشارات الصوتية.

تم الاعتماد على مصفوفة طاقة المرشحات (Filter Bank Energies) لاستخراج الميزات الصوتية بدلاً من الطرق التقليدية مثلMFCC ، حيث توفر هذه الطريقة تمثيلاً أكثر تعبيراً ودقة للطيف الصوتي. كما تم تطبيق تقنية SpecAugmentوهي إحدى تقنيات Data Augmentation المتقدمة، والتي تعتمد على إخفاء أجزاء من طيف الصوت زمنياً أو ترددياً، مما يساعد النموذج على التعميم والتعامل مع تنوع البيانات الفعلية.

استخدم البحث خوارزمية Connectionist Temporal Classification (CTC) لتعيين تسلسل الإدخال (الإشارات الصوتية) إلى تسلسل الإخراج (النص)، حيث توفر هذه الخوارزمية إمكانية التعامل مع تسلسلات ذات أطوال غير متساوية دون الحاجة إلى محاذاة دقيقة.

كما تم تدريب النموذج باستخدام تقنية الانتشار العكسي (Backpropagation) لتحديث الأوزان بناءً على دالة الخسارة الناتجة عن CTC، وتم تقييم أداء النموذج من خلال معدل خطأ الكلمات (Word Error Rate – WER)، وهو المعيار الشائع لتقييم دقة أنظمة التعرف التلقائي على الكلام.

٦- مراحل البحث

٦-١ مرحلة الحصول على البيانات:

تشكل البيانات الصوتية أحد الركائز الأساسية لتطوير نموذج قوي وفعّال في التعرف التلقائي على الكلام. يعتمد البحث على تنوع في مصادر البيانات الصوتية لضمان شمولية النموذج وتعميمه على لهجات وسيناريوهات متعددة. يساهم دمج مجموعات بيانات متنوعة في إثراء النموذج بأنماط نطق مختلفة، وجودة تسجيل متفاوتة، وسياقات لغوية متعددة، مما يعزز من دقته وموثوقيته عند تطبيقه في بيئات واقعية. وقد تم استخدام ثلاث مجموعات بيانات أساسية في هذا البحث:

٦-١-١ The LJ Speech Dataset :

مجموعة بيانات مفتوحة المصدر تتضمن حوالي ١٣,١٠٠ مقطع صوتي لقراءة مقاطع من كتب غير خيالية من قِبل متحدث واحد (أنثى). يتراوح طول المقاطع بين ثانية واحدة إلى عشر ثوانٍ، بإجمالي يقارب ٢٤ ساعة صوتية، بحجم يصل إلى ٢,٦ جيجابايت. تم تسجيل هذه المقاطع ضمن مشروع LibriVox وتتوفر مع النصوص المقروءة [10].

٦-١-٢ : CREMA-Dataset :

تحتوي على ٧,٤٤٢ مقطعاً صوتياً صُوّر من قبل ٩١ ممثلاً وممثلة بأعمار وخلفيات عرقية متنوعة، ما يمنح البيانات تنوعاً غنياً في اللهجات والتعبير الصوتي. تستخدم هذه البيانات بشكل واسع في أبحاث تحليل العاطفة والتعرف على النطق، وتُعد مناسبة لتقييم قدرة النموذج على التكيف مع أنماط تعبيرية مختلفة [11].

٦-١-٣ LibriSpeech ASR :

تعد من أشهر مجموعات البيانات في مجال ASR، حيث تحتوي على أكثر من ١٠٠ ساعة من الكتب الصوتية المقروءة باللغة الإنجليزية، مسجلة بجودة ١٦kHz. تم تقسيم البيانات بعناية إلى مجموعات تدريب واختبار لضمان الاتساق والموثوقية في التجارب. الحجم الكلي للبيانات يبلغ حوالي ٦,٣ جيجابايت[12].

٦-٢ الطريقة المقترحة

تم في هذا البحث اقتراح نموذج متقدم للتعرف التلقائي على الكلام قائم على معماريةConformer ، وهي بنية هجينة تدمج بين الشبكات التلافيفية وآليات الانتباه الذاتي(Self-Attention) ، مما يسمح للنموذج بالتقاط الأنماط الزمنية الدقيقة والمعلومات السياقية بعيدة المدى بشكل متوازن وفعّال.

اعتمد النموذج على مجموعة من الخطوات المنهجية، أبرزها استخدام تقنيات زيادة البيانات (Data Augmentation) باستخدامSpecAugment ، التي توفر تنوعاً كبيراً في مدخلات النموذج، حيث يتم تنفيذ تغييرات مخططة على البيانات الأصلية مثل إخفاء نطاقات زمنية أو ترددية من الطيف الصوتي. وتُعد هذه الخطوة من أهم الإسهامات في تحسين قدرة النموذج على التعميم على بيانات حقيقية تتضمن لهجات مختلفة وضوضاء بيئية متنوعة، وهي نقطة لم تُعالج بشكل كافٍ في الدراسات السابقة.

كما تم استخدام تقنية Filter Bank Energies لاستخراج الميزات الصوتية من الإشارات الخام، وهي تقنية أكثر تعبيراً مقارنة بـ MFCC، وتم تحسين هذه الميزات عبر التطبيع (Feature Normalization)، لضمان الاتساق بين المقاطع الصوتية المختلفة وتقليل تأثير اختلافات التسجيل. خطوات بناء النموذج المقترح سنقوم بتلخيصها كما يلي:

١- المعالجة الأولية للمقاطع الصوتية (Pre-processing)

٢- زيادة البيانات باستخدام SpecAugment

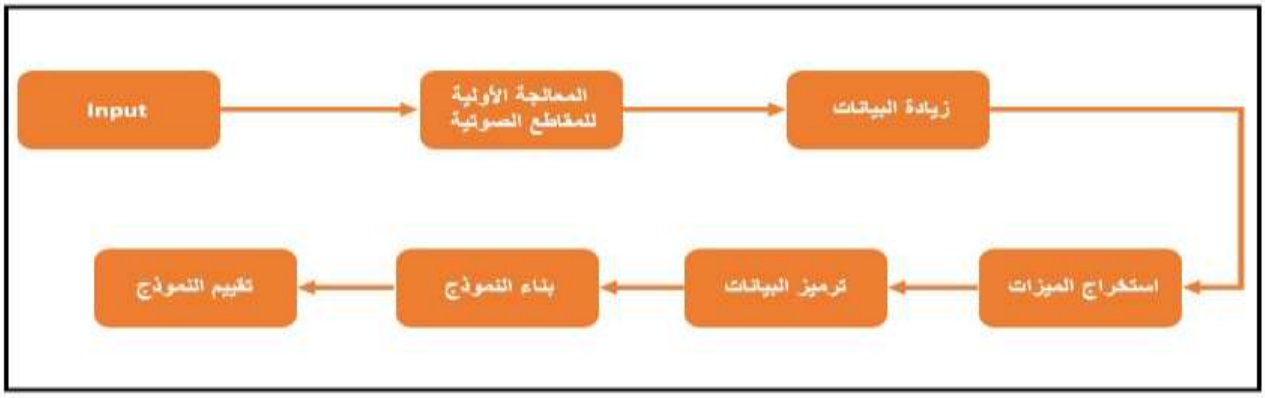
٣- استخراج الميزات (Feature Extraction)

٤- ترميز البيانات الصوتية والنصية

٥- بناء النموذج باستخدام معمارية Conformer

٦- تقييم النموذج

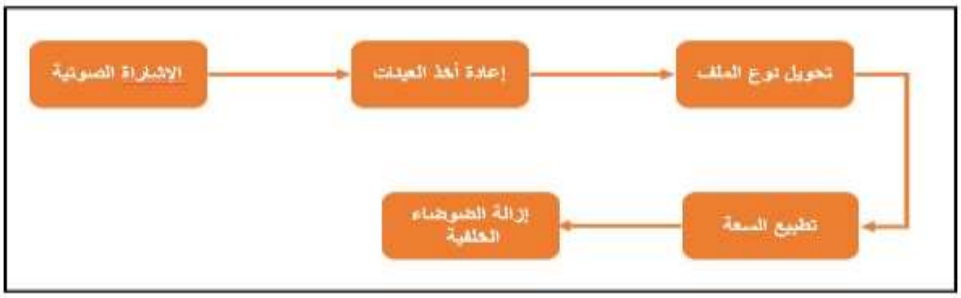
كما يوضح الشكل (1)، ويتم شرح الطريقة المقترحة بشكل تفصيلي في الفقرات التالية من البحث.



الشكل (1): الخطوات المتبعة لبناء النموذج المقترح

٦-2-١: المعالجة الأولية للمقاطع الصوتية Pre-Processing:

تُعد المعالجة الأولية للمقاطع الصوتية خطوة محورية وأساسية في سلسلة بناء أنظمة التعرف التلقائي على الكلام [13]، حيث تهدف إلى تهيئة الإشارات الصوتية الخام لتكون قابلة للاستخدام في مراحل لاحقة مثل استخراج الميزات والتعلم الآلي. في هذا البحث، تم اعتماد إجراءات دقيقة لضمان جودة واستقرار البيانات الصوتية، وذلك وفقاً للخطوات التالية كما هي موضحة بالشكل (2):



الشكل (2): خطوات المعالجة الأولية للمقاطع الصوتية

١- إعادة أخذ العينات Resampling: تم توحيد معدل أخذ العينات لجميع المقاطع الصوتية إلى ١٦ كيلوهرتز [14]، وهو المعدل المستخدم بشكل شائع في تطبيقات التعرف التلقائي على الكلام لتقليل حجم البيانات مع الحفاظ على جودة الصوت.

٢- تحويل نوع الملفAudio Format Conversion : تم تحويل كافة الملفات الصوتية إلى تنسيق WAV أحادي القناة (Mono) لتسهيل عمليات المعالجة وضمان التوافق مع خوارزميات استخراج الميزات.

٣- تطبيع السعة Amplitude Normalization : تم ضبط سعة الإشارات الصوتية إلى نطاق موحد يتراوح بين −١ و +١ لضمان ثبات المستويات الصوتية وتحقيق توافق أفضل عند المعالجة باستخدام الشبكات العصبية.

٤- إزالة الضوضاء الخلفيةNoise Reduction : تم تطبيق خوارزميات تصفية بسيطة لتحسين نقاء الصوت، وذلك دون التأثير على محتوى الكلام الفعلي.

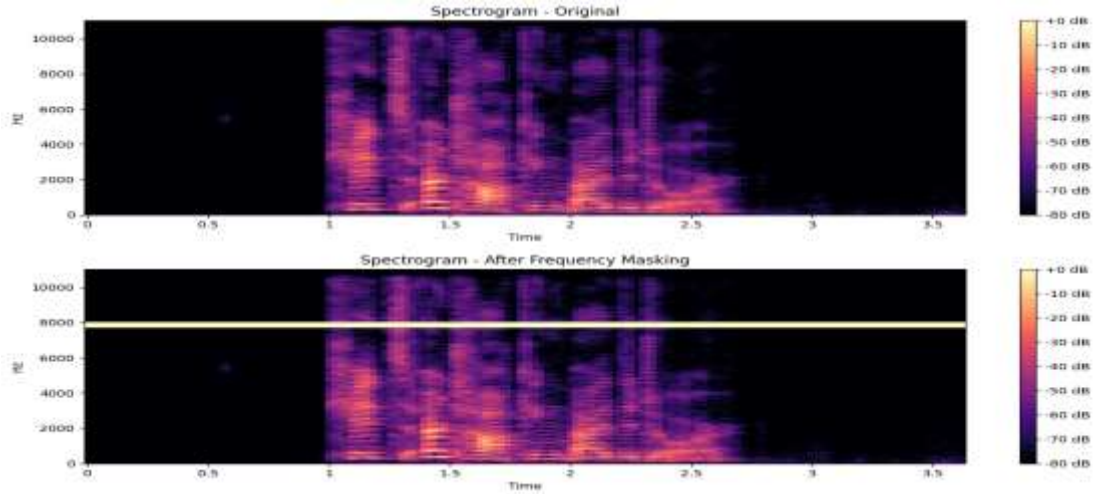
تساهم هذه الخطوات في إعداد المقاطع الصوتية بشكل مناسب للنموذج المقترح، مما يعزز من دقة عمليات استخراج الميزات والتدريب، ويسهم في تحقيق أداء أعلى في تحويل الكلام إلى نص. وتُعد هذه المعالجة أساساً لأي نظام دقيق وموثوق في مجال التعرف التلقائي على الكلام.

٦-٢-٢ زيادة البيانات باستخدام SpecAugment :

تُعد زيادة البيانات(Data Augmentation) من أهم الاستراتيجيات المستخدمة لتحسين أداء نماذج التعلم العميق في مهام التعرف التلقائي على الكلام، حيث تُسهم في تعزيز قدرة النموذج على التعميم ومقاومة الضوضاء وتنوع اللهجات والظروف الصوتية المختلفة. في هذا البحث، تم استخدام تقنية SpecAugment وهي من أحدث تقنيات زيادة البيانات المعتمدة في أنظمة تحويل الكلام إلى نص، وتُطبق مباشرة على تمثيل الطيف الصوتي Spectrogram.

تعتمد SpecAugment على ثلاث مكونات رئيسية لزيادة التنوع في بيانات التدريب:

١- **إخفاء التردد (Frequency Masking):** تُعد تقنية إخفاء التردد (Frequency Masking) أحد أساليب زيادة البيانات (Data Augmentation) الفعالة التي أثبتت جدواها في تدريب نماذج التعرف التلقائي على الكلام (ASR). تعتمد هذه التقنية على حجب نطاقات ترددية معينة من طيف الإشارة الصوتية (Spectrogram) بشكل عشوائي خلال مرحلة التدريب، وذلك بهدف تحسين قدرة النموذج على التعميم والتعامل مع فقدان المعلومات الترددية في البيئات الواقعية، مثل الضوضاء أو ضعف جودة التسجيل.



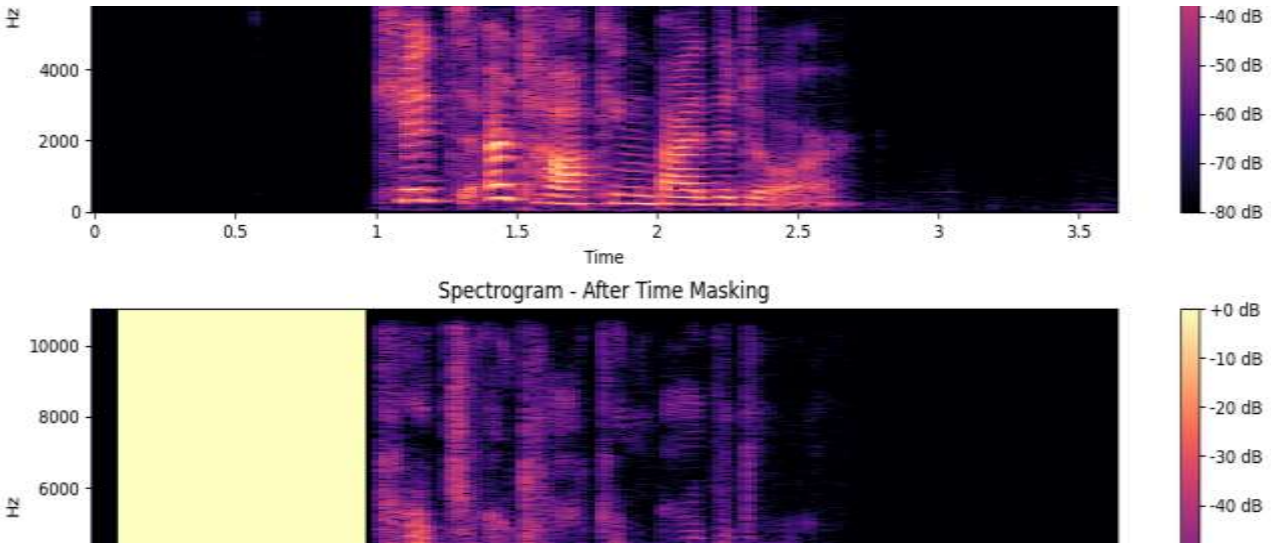
الشكل (3): تغير شكل الإشارة الصوتية بعد إخفاء التردد

عند تطبيق إخفاء التردد، يتم اختيار عدد من الشرائط الترددية ضمن طيف الإشارة (عادة بشكل مستطيل عمودي على محور الزمن) وتُستبدل بقيم صفرية أو عشوائية. هذا الإجراء يُشبه "تشويشًا اصطناعيًا" يُجبر النموذج على التركيز على السياق الزمني والمعلومات التكميلية المتبقية في الإشارة، دون الاعتماد فقط على ترددات معينة، مما يُعزز من مرونته ويقلل من خطر التعميم المفرط (Overfitting).

في هذه الدراسة، قمنا بتطبيق تجربة على مقطع صوتي واحد، حيث تم تصوير الطيف الترددي للإشارة قبل الإخفاء، فظهر بشكل طبيعي يتضمن جميع النطاقات الترددية. بعد تطبيق تقنية الإخفاء، لاحظنا ظهور شرائط ترددية داكنة تشير إلى المناطق التي تم حجبها. ورغم هذا الحجب الجزئي، بقيت الإشارة مفهومة إلى حد كبير، مما يدل على أن النموذج سيظل قادرًا على فهم السياق واستخلاص المعنى حتى عند فقدان بعض المعلومات.

تُعتبر هذه التقنية جزءًا من خوارزمية SpecAugment، وقد أثبتت الدراسات أنها تساهم بشكل كبير في تحسين أداء نماذج Conformer وغيرها من هياكل ASR، خاصة في بيئات بها تشويش، لهجات مختلفة، أو تسجيلات منخفضة الجودة. كما تُعد مناسبة جدًا للتدريب على بيانات محدودة، لأنها تضيف تنوعًا زائفًا يساعد على بناء نموذج أكثر استقرارًا وقابلية للتعميم.

٢- **إخفاء الزمن (Time Masking):** تُعد تقنية إخفاء الزمن (Time Masking) إحدى تقنيات زيادة البيانات الصوتية (Audio Data Augmentation) المستخدمة في تحسين كفاءة ودقة نماذج التعلم العميق، وخصوصًا في أنظمة التعرف التلقائي على الكلام (ASR). تقوم هذه التقنية على حجب (إخفاء) مقاطع زمنية متعاقبة من طيف الإشارة الصوتية (Spectrogram)، بحيث تُزال بشكل عشوائي شرائط على محور الزمن. هذا الإخفاء يُجبر النموذج على التركيز على المعلومات السياقية والزمنية الأوسع بدلاً من الاعتماد على تفاصيل لحظية قد



الشكل (4): تغيير مدة الإشارة الصوتية بعد إخفاء الزمن

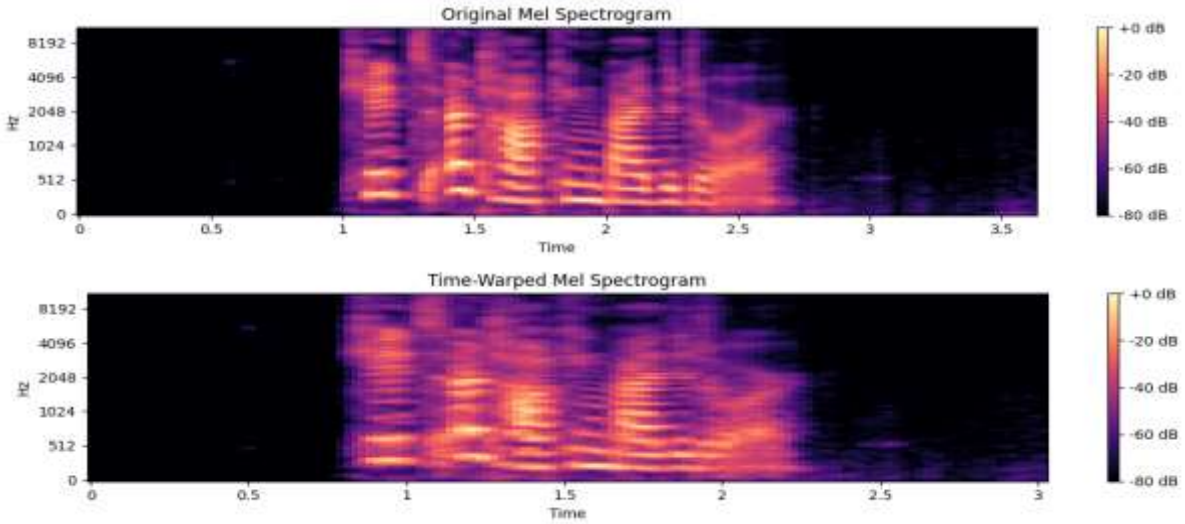
تتعرض للضياع أو التشويش في البيئات الواقعية.

تكمن الفكرة الرئيسة من هذه التقنية في محاكاة الانقطاعات أو التشويش الزمني الذي قد يحدث في التسجيلات الفعلية – مثل فقدان جزء من الكلمة، أو وجود صمت مفاجئ أو ضجيج بيئي – مما يجعل النموذج أكثر مرونة وقدرة على التنبؤ الصحيح حتى في ظروف غير مثالية.

في هذه الدراسة، قمنا بتطبيق تجربة على مقطع صوتي باستخدام تقنية Time Masking. أظهر الطيف الزمني للإشارة الصوتية قبل المعالجة تمثيلًا كاملًا للمعلومات الزمنية. وبعد التطبيق، ظهرت شرائط عمودية معتمة في الطيف، تدل على الفترات الزمنية التي تم إخفاؤها عمدًا. ورغم هذا الإخفاء، ظل النموذج قادرًا على فهم السياق العام للإشارة، مما يعكس قدرته على استخلاص المعنى من بيانات غير مكتملة.

تُستخدم هذه التقنية عادةً كجزء من خوارزمية SpecAugment، وهي تُعزز من الاستقرار والتنوع في بيانات التدريب، وتُساعد على الحد من التعميم المفرط للنموذج. كما أن دمجها مع تقنيات أخرى مثل إخفاء التردد يساهم في تحسين أداء النماذج على مختلف مجموعات البيانات، بما في ذلك تلك المسجلة في ظروف صعبة أو ذات جودة منخفضة.

٣- **التمدد الزمني(Time Warping)** : تُعد تقنية التمدد الزمني (Time Warping) من الركائز الأساسية في خوارزمية SpecAugment، وهي إحدى أهم تقنيات زيادة البيانات الصوتية (Data Augmentation) المستخدمة في تحسين أداء أنظمة التعرف التلقائي على الكلام (ASR). تعتمد هذه التقنية على تشويه غير



الشكل (5): إزاحة الإشارة الصوتية من خلال التمدد الزمني

خطي للطيف الصوتي من خلال إزاحة مكونات معينة منه على طول محور الزمن، بطريقة تُحاكي التغيرات الديناميكية التي تطرأ على سرعة أو إيقاع النطق بين المتحدثين.

الهدف الأساسي من هذه التقنية هو تعريض النموذج لسيناريوهات متعددة لأنماط النطق، حيث تختلف السرعة، وتتنوع التوقفات والانتقالات بين المقاطع الصوتية من شخص لآخر. من خلال هذا التشويه الزمني، يصبح النموذج أكثر مرونة في التعامل مع التباينات الطبيعية في التحدث، مثل التباطؤ أو التسارع في الكلام، والتي كثيرًا ما تواجهها أنظمة ASR في التطبيقات الواقعية.

في إطار التجارب العملية التي أُجريت في هذه الدراسة، تم تطبيق تقنية Time Warping باستخدام خوارزمية تُحدث انحرافًا محليًا في طيف الإشارة الصوتية، بحيث يتم اختيار نقطة عشوائية داخل الطيف الصوتي ثم إزاحتها مؤقتًا باستخدام منحنى انحناء زمني. أظهرت مقارنة الطيف قبل وبعد التطبيق أن الطيف الأصلي كان منتظمًا ومتناسقًا

زمنيًا، في حين ظهر بعد المعالجة انحراف ملحوظ في بعض المقاطع، مما يعكس حدوث "تمدّد زمني" اصطناعي يُعزز من واقعية البيانات المقدمة للنموذج.

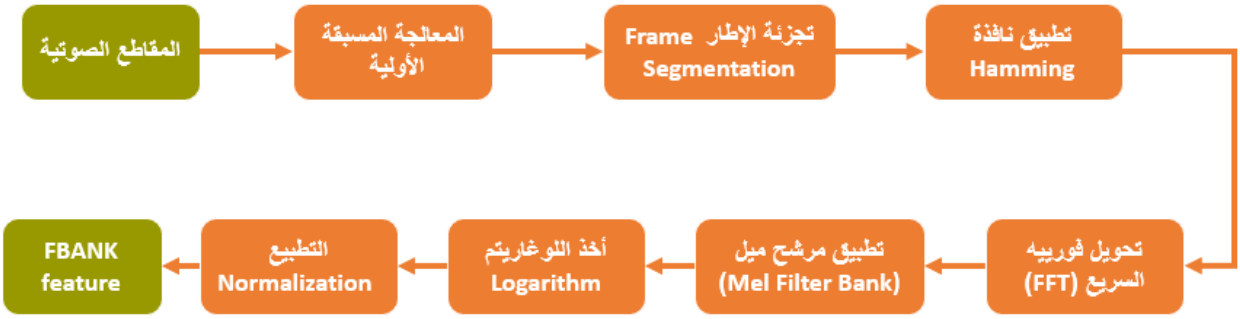
وتُسهم هذه التقنية في رفع قدرة النموذج على التعميم، حيث يتعلم النموذج التعامل مع تسجيلات تحتوي على تفاوت زمني طبيعي أو غير منتظم. كما أن دمجها مع تقنيات مثل إخفاء التردد وإخفاء الزمن ضمن إطار SpecAugment، يؤدي إلى إنتاج نموذج أكثر استقرارًا وفعالية، وذو قدرة أكبر على التعرف على الكلام في بيئات متنوعة ومتغيرة.

٦-٤-٣ : استخراج الميزات Feature Extraction : يمكن تعريف استخراج الميزات بأنها عملية تحويل الإشارة الصوتية الخام إلى تمثيل رقمي مختصر يحتفظ بالمعلومات الجوهرية ذات الصلة بالتحليل أو التنبؤ. وتُعد هذه الخطوة ضرورية في أنظمة التعرف التلقائي على الكلام، حيث تسهم في تقليل حجم البيانات مع الاحتفاظ بالمكونات المهمة التي تساعد النموذج على التمييز بين الأصوات أو الكلمات المنطوقة.

في هذه الدراسة، تم استخدام تقنية طاقات مرشحات بنك التردد (Filter Bank Energies – FBANK)، والتي تعد من الأساليب الشائعة لاستخراج الميزات من الصوت [15]، وتشكل بديلاً مباشراً لتقنية MFCC، إذ تعتمد على نفس البنية الأساسية لكن دون تطبيق تحويل جيب التمام المنفصل (DCT).

تُظهر ميزات FBANK الطيف الطاقي لإشارة الصوت على مقياس الإدراك البشري للتردد (Mel scale)، مما يجعلها حساسة للخصائص الطيفية الدقيقة للصوت، وتسمح للنموذج بالتعلم مباشرة من تلك الأنماط.

يوضح الشكل (6) خطوات استخراج الميزات باستخدام FBANK :



الشكل (6): خطوات استخراج الميزات باستخدام FBANK

المعالجة المسبقة (Pre-processing): يتم أخذ عينات من الإشارة الصوتية بمعدل ثابت (16 كيلوهرتز) لتحويل الإشارة التناظرية إلى رقمية قابلة للمعالجة. بعد ذلك، تُطبّق عملية التطبيع لضبط سعة الإشارة بحيث تقع في نطاق قياسي (-١ إلى ١) مما يساعد على استقرار النموذج أثناء التدريب.

تجزئة الإطار (Frame Segmentation): تُقسّم الإشارة الصوتية إلى إطارات زمنية قصيرة (بين ٢٠ إلى ٥٠ مللي ثانية) لأن خصائص الصوت تتغير بمرور الوقت. يتيح هذا التقسيم تحليل الإشارة ضمن نوافذ صغيرة يمكن اعتبارها ثابتة إحصائياً، مما يسهل استخلاص ميزات دقيقة.

تطبيق نافذة Hamming (Windowing): يتم تطبيق دالة نافذة من نوع Hamming على كل إطار لتقليل التأثيرات الحادة عند حواف الإطار.يساعد ذلك في تقليل ظاهرة التسرب الطيفي (Spectral Leakage) التي قد تشوش على تمثيل الترددات في الإشارة.

تحويل فورييه السريع (Fast Fourier Transform - FFT): تُستخدم خوارزمية FFT لتحويل كل إطار من المجال الزمني إلى المجال الترددي، مما يسمح بتحليل المكونات الترددية. هذا التحويل ضروري لفهم توزيع الطاقة في ترددات الإشارة، وهي خطوة جوهرية في استخراج الميزات.

تطبيق مرشح ميل (Mel Filter Bank): يتكوّن بنك مرشحات Mel من مجموعة مرشحات مثلثية يتم توزيعها بشكل غير خطي لتتناسب مع مقياس Mel الإدراكي. يهدف هذا إلى محاكاة الطريقة التي تدرك بها الأذن البشرية الترددات، حيث تكون أكثر حساسية للترددات المنخفضة.

أخذ اللوغاريتم (Logarithm): يتم حساب اللوغاريتم لطاقات المرشحات الناتجة من بنك Mel لتحويل التمثيل الخطي للطاقة إلى تمثيل لوغاريتمي. هذا التمثيل أقرب لطريقة إدراك البشر لشدة الصوت، كما يُبرز الفروقات الدقيقة في الطيف الترددي.

التطبيع (Normalization): تُطبّق عملية تطبيع إحصائي على متجهات الميزات (z-score) لجعل البيانات أكثر اتساقاً. يساعد هذا في تقليل تأثير الاختلافات الفردية في التسجيلات مثل اختلاف المسافات أو الضوضاء الخلفية.

ميزات FBANK (FBANK Features): بعد تطبيق الخطوات السابقة، نحصل على مجموعة من المتجهات التي تمثل الطيف الطاقي على مقياس Mel بشكل مباشر. تُستخدم هذه الميزات كمدخلات فعّالة لنماذج التعلم العميق في مهام مثل التعرف التلقائي على الكلام.

٦-٢-٤ ترميز البيانات (الصوت والنص):

في أنظمة التعرف التلقائي على الكلام (ASR)، يُعد ترميز البيانات خطوة جوهرية تهدف إلى تحويل الإشارات الصوتية والنصوص المصاحبة لها إلى تمثيل رقمي يمكن للنموذج التعلمي التعامل معه بفعالية خلال مرحلة التدريب.

ترميز الصوت (Audio Encoding): يتم استخراج الميزات الصوتية باستخدام تقنيات FBANK التي تكلمنا عنها سابقا، حيث تُحوَّل الموجة الصوتية إلى تمثيل رقمي يعكس المعلومات الترددية للإشارة الصوتية. هذه الميزات تُستخدم كمدخلات للنموذج أثناء التدريب.

ترميز النص (Text Encoding): يتم تحويل كل عنصر نصي (مثل الحروف أو الرموز) إلى قيمة رقمية باستخدام تقنيات مثل tokenizer. يتيح هذا الترميز للنموذج الربط بين الميزات الصوتية والنصوص الفعلية المنطوقة بهدف التعلم وتحقيق التطابق.

٦-2-٥ بناء الموديل باستخدام Conformer :

في السنوات الأخيرة، أصبحت بنية Conformer واحدة من أقوى وأكثر البنى تطورًا في مجال التعرف التلقائي على الكلام (ASR)، حيث تمكنت من تحقيق نتائج رائدة في العديد من التحديات العالمية. يجمع هذا النموذج بين مزايا الطبقات التلافيفية (CNN) والطبقات التحويلية (Transformer) في تصميم معماري واحد متكامل، مما يجعله مناسبًا بشكل خاص للتعامل مع الإشارات الصوتية المعقدة والغير متزامنة. تم تطوير Conformer خصيصاً للتغلب على نقاط الضعف في النماذج السابقة مثل RNN وCRNN من خلال دمج آليات الانتباه مع المعالجة المحلية.

٦-٢-٥-١ بنية النموذج Model Architecture:

يتألف نموذج Conformer من ثلاث مكونات أساسية تعمل بتناغم لاستخلاص التمثيلات الغنية من الإشارات الصوتية:

- الطبقات التلافيفية الأولية (Initial Convolutional Layers) : تمر البيانات أولاً عبر طبقات تلافيفية ثنائية الأبعاد، تهدف إلى استخراج الأنماط الترددية والإيقاعية المحلية. هذا المعالجة الأولية تساعد في تقليل التشويش وتحسين جودة الخصائص الصوتية المستخلصة، مما يمهد الطريق لمرحلة تحليل أعمق وأكثر تعقيداً.
- طبقات (Conformer (Conformer Blocks: تمثل هذه الطبقات جوهر النموذج، وهي السبب الرئيسي وراء تفوق Conformer على البنى التقليدية. كل طبقة Conformer تتضمن أربع وحدات أساسية:
- وحدة Feed-Forward Network (FFN): تقوم بتوسيع تمثيل البيانات وتحليله من خلال شبكة عصبية كثيفة، مما يساعد النموذج على اكتساب تمثيل موزع وعام للخصائص الصوتية.
- وحدة Multi-Head Self-Attention (MHSA) : تتيح للنموذج استيعاب العلاقات البعيدة بين المقاطع الصوتية، وهو أمر بالغ الأهمية لفهم السياق الكامل للكلام، خاصة في الجمل الطويلة.
- وحدة Convolution Module : تضيف قدرة النموذج على التقاط الخصائص المحلية الدقيقة (مثل تغير الترددات بسرعة)، وتعمل كمرشح لإزالة الضوضاء وتحسين دقة التنبؤ.
- الروابط المتبقية (Residual Connections) والتطبيع (Layer Normalization) : تساهم هذه الآليات في استقرار عملية التدريب وتقليل التذبذبات، مما يسمح بتدريب أعمق دون فقدان المعلومات أو ظهور مشكلة انفجار التدرجات.
- قوة Conformer تكمن في الجمع الذكي بين السياق العالمي (Global Context) من خلال Attention، والمعالجة المحلية (Local Feature Extraction) عبر Convolution
- الطبقات الكثيفة (Dense Output Layers) : بعد المرور بسلسلة طبقاتConformer ، يتم تحويل المخرجات إلى تمثيل ثنائي الأبعاد باستخدام طبقات كثيفة. ثم يتم تطبيق دالة تنشيط Softmax على المخرجات، لإنتاج توزيع احتمالي على جميع الرموز والأحرف ضمن المفردات المستهدفة، مما يسمح باختيار التنبؤ الأنسب في كل خطوة زمنية.

٦-٢-٥-٢ تجميع النموذج Model Compilation :

المُحسّن Optimizer: يتم استخدام المحسن Adam بفضل سرعته وقدرته على التكيف، مع معدل تعلم مضبوط لتوفير توازن جيد بين سرعة التدريب والاستقرار.

دالة الخطأ Loss Function: يتم اعتماد CTC Loss (Connectionist Temporal Classification)، والتي تُعد مثالية لمهام التعرف على الكلام حيث لا يتطابق طول تسلسل الإدخال (الإشارة الصوتية) مع تسلسل الإخراج (النص).

تتيح هذه الدالة للنموذج تعلم المحاذاة المثلى بين الإشارات الصوتية والتسميات النصية دون الحاجة لتحديد مواقع الأحرف بدقة.

٦-٢-٥-٣ معدل خطأ الكلمات (Word Error Rate (WER:

يُعد معدل خطأ الكلمات (WER) أحد أهم المقاييس القياسية المستخدمة لتقييم دقة أنظمة التعرف التلقائي على الكلام (ASR). يقيس هذا المؤشر نسبة الكلمات التي أخطأ النموذج في التعرف عليها مقارنةً بالنص المرجعي الصحيح كما هو مشار اليه في المعادلة (١). كلما كان WER أقل، دلّ ذلك على أداء أقوى وأكثر دقة للنظام. يعتمد هذا المقياس على مقارنة تسلسل الكلمات في الإخراج النصي الذي تولده خوارزمية التعرف، مع التسلسل الصحيح في المرجع اليدوي، ويتم حسابه باستخدام الصيغة التالية[16]:

$$WER = \frac{S + D + I}{N} \qquad (1)$$

حيث:

S: عدد البدائل (Substitutions) – الكلمات التي استبدلها النظام بكلمات خاطئة.

D: عدد عمليات الحذف (Deletions) – الكلمات التي لم يتعرف عليها النظام إطلاقاً.

I: عدد الإدخالات (Insertions) – الكلمات التي أضافها النظام بالخطأ.

N: إجمالي عدد الكلمات في النص المرجعي.

يتم احتساب WER عن طريق محاذاة النص الناتج مع النص المرجعي كلمةً بكلمة، وتحديد نوع الخطأ الحاصل في كل حالة، أنواع الأخطاء:

١) البدائل: استبدال كلمة صحيحة بكلمة خاطئة.

٢) الإدخالات: إضافة كلمة غير موجودة في النص الأصلي.

٣) الحذف: حذف كلمة من النص المرجعي وعدم التعرف عليها.

كلما اقترب WER من الصفر، دلّ ذلك على أن النظام يقترب من الأداء المثالي، إذ تعني WER بنسبة ٠% أن جميع الكلمات في النص المرجعي تم التعرف عليها بشكل صحيح دون أي أخطاء.

٧- النتائج والمناقشة:

يستعرض هذا الجزء من المقالة نتائج معالجة الإشارات الصوتية واستخراج السمات باستخدام تقنية طاقات مرشحات بنك التردد (*Filter Bank Energies - FBANK*)، إضافة إلى نتائج نموذج *Conformer* في مهمة التعرف على الكلام. تم تنفيذ مراحل التدريب والتقييم باستخدام بيئة *Python* (الإصدار ٣,٨,٣)، مع الاعتماد على مكتبات متقدمة في معالجة الصوت والتعلم العميق مثل *TensorFlow* و *Librosa*.

٧-١ سمات مجموعة البيانات:

تم اعتماد ٥٩ سمة صوتية مستخرجة باستخدام تقنية طاقات مرشحات بنك الترددFBANK - Filter Bank Energies)* ، كما هو موضح في الجدول (١)، وهي تتيح تمثيلاً طيفياً دقيقاً للإشارة الصوتية الخام دون اللجوء إلى إسقاطات مثل تحويل *DCT* المستخدم في *MFCC* تحتفظ هذه السمات بالخصائص الزمنية والترددية الدقيقة للإشارة، مما يجعلها مناسبة جداً لنماذج التعلم العميق مثل*Conformer* تشمل السمات المستخرجة خصائص متعددة مثل الطاقة الطيفية لكل فلتر، تدفق الإشارة عبر الزمن، والتباين الطيفي بين النوافذ الزمنية، مما يوفر تمثيلاً غنياً يساعد على التعرف على الكلمات بدقة عالية، حتى في البيئات المليئة بالضوضاء أو عند وجود لهجات مختلفة.

جدول (١) : سمات مجموعة البيانات

وصف السمات	السمات
طاقة الإشارة في نطاق التردد الأول (أقل الترددات)	FBANK 1
طاقات الفلاتر البنكية في النطاقات المختلفة (تغطي الترددات بين ٠-٨٠٠٠ *Hz*)	FBANK 2➔ 43
المشتقات الزمنية الأولى (*delta*) لقياس التغير عبر الزمن	Δ FBANK 43 ➔ 50
معاملات الترتيب الأعلى تلتقط التفاصيل الدقيقة	Δ Δ FBANK 51 ➔ 58
الطاقة الإجمالية :متوسط الطاقة الكلي لكل إطار صوتي	FBANK 59

٧-٢ تقليل السمات:

يشير تقليل السمات في سياق التعلم العميق إلى عملية اختيار مجموعة فرعية من السمات الأكثر ارتباطاً بالفئة المستهدفة أو الأقل تكراراً للمعلومات، بهدف تحسين أداء النموذج وتقليل التعقيد الحسابي. إن وجود سمات زائدة أو متكررة قد يؤدي إلى زيادة زمن الحساب دون تحسين فعلي في الأداء، بل قد يسبب في بعض الحالات تدهور دقة التصنيف. لذا تُعدّ هذه العملية خطوة ضرورية.

في هذا البحث، تم الاعتماد على تحليل معامل الارتباط كوسيلة فعالة لتحديد السمات غير المفيدة، حيث تتيح هذه الطريقة الكشف عن السمات التي ترتبط بشدة بسمات أخرى، مما يشير إلى احتمال وجود معلومات مكررة يمكن الاستغناء عنها.

٧-٢-١ تحليل معامل الارتباط:

معامل الارتباط هو مقياس إحصائي يعبر عن قوة العلاقة بين سميتين، ويتراوح بين −١ و +١. كلما اقتربت القيمة من +١ أو −١، دلّ ذلك على وجود علاقة قوية. في حال وجود ارتباط قوي جداً (مثلاً > ٠,٩٥ أو < −٠,٩٥)، يتم الاحتفاظ بإحدى السمتين وحذف الأخرى لتجنب التكرار.

الصيغة الرياضية لحساب معامل الارتباط[16] :

$$r = \frac{\sum(xi - \bar{x})(yi - \bar{y})}{\sqrt{\sum(xi - \bar{x})^2 \sum(yi - \bar{y})^2}} \qquad (2)$$

حيث:

xi , yi : القيم الفردية للسمتين قيد التحليل.

$\bar{x}$, $\bar{y}$: المتوسط الحسابي لكل سمة.

بالاعتماد على العلاقة السابقة تم تنفيذ هذه الخطوة باستخدام مكتبات pandas وnumpy في لغة Python، حيث تم حساب مصفوفة الارتباط لجميع السمات، ثم استخراج الأزواج التي أظهرت ارتباطاً مرتفعاً. فكانت النتيجة كما هو موضح بالجدول (٢):

جدول (٢) : أزواج السمات المرتبطة

السمة الأولى	السمة الثانية	معامل الارتباط	القرار
FBANK(31)	FBANK(33)	0.97	حذف FBANK (31)
FBANK(37)	FBANK(35)	−0.96	حذف FBANK (37)
FBANK(44)	FBANK(42)	0.95	حذف FBANK (44)
FBANK(48)	FBANK(47)	0.96	حذف FBANK (48)
FBANK(53)	FBANK(55)	0.96	حذف FBANK (53)
FBANK(59)	FBANK(57)	0.95	حذف FBANK (59)

بناءً على هذا التحليل، تم تحديد ٦ سمات مرتبطة ارتباطاً عالياً بسمات أخرى، وتم حذفها وبذلك، انخفض عدد السمات من ٥٩ إلى ٥٣ سمة فقط، مما ساعد على تقليل حجم البيانات وتحسين كفاءة النموذج دون التأثير على دقة التصنيف.

٣-٧ بيانات التدريب وبيانات التحقق:

تم تدريب النموذج باستخدام مجموعة بيانات صوتية تتضمن تسجيلات تمثل كلمات منطوقة بصيغ مختلفة. شكّلت هذه البيانات الأساس الذي يُمكِّن النموذج من تعلم الأنماط الصوتية واستخلاص السمات التمييزية المرتبطة بكل فئة.

وبهدف ضمان قدرة النموذج على التعميم وعدم الانحياز لبيانات التدريب فقط، تم تخصيص جزء منفصل من البيانات لاستخدامه كمجموعة للتحقق (Validation Set) تم تقييم أداء النموذج على هذه المجموعة في نهاية كل دورة تدريبية (Epoch)باستخدام وظيفة رد نداء (Callback) مخصصة.

تتضمن هذه الوظيفة حساب معدل خطأ الكلمات(Word Error Rate - WER) ، وهو مقياس يعكس دقة التعرف الصوتي من خلال مقارنة المخرجات المتوقعة بالنصوص الأصلية الفعلية. إضافةً إلى ذلك، يتم عرض عينات من التنبؤات بغرض التقييم النوعي لأداء النموذج وملاحظة أي حالات فشل أو تحسن ملحوظة مع مرور الوقت .ساهمت هذه

الآلية في مراقبة الأداء وتحديد النقطة المثلى لإيقاف التدريب المبكر (Early Stopping) في حال ملاحظة انخفاض في دقة التحقق، مما ساعد على تحسين فعالية النموذج النهائي.

٧-٤ تقييم أداء النموذج:

تم تدريب النموذج لمدة ٥٠ حقبة زمنية باستخدام وحدة معالجة الرسومات GeForce RTX 2080 Ti، حيث استغرقت كل حقبة ما يقرب من 22 إلى 26 دقيقة ما يقارب ٢٠ ساعة. وقد أظهرت النتائج تحسنًا ملحوظًا في أداء النموذج مع تقدم التدريب، وذلك بعد تطبيق تقنية اختزال السمات إلى ٥٣ سمة ذات ارتباط أعلى بأهداف التعرف على الكلمات كما هو موضح في الجدول (٣):

جدول (٣) :معدل الخطأ الكلمات والخسارة خلال كل حقبة

Epoch	WER	Val-loss	Epoch	WER	Val-loss
26	0.5054	80.4710	**1**	**0.9999**	**270.8235**
27	0.4837	72.8569	**2**	**0.9800**	**263.2094**
28	0.4628	65.2428	3	0.9643	255.5953
29	0.4447	57.6287	4	0.9426	247.9812
30	0.4238	50.0146	5	0.9200	240.3671
31	0.3901	44.7862	6	0.9033	232.7530
32	0.3717	39.5578	7	0.8814	225.1389
33	0.3517	34.3294	8	0.8665	217.5248
34	0.3337	29.1010	9	0.8434	209.9107
35	0.3136	26.9834	10	0.8219	202.2966
36	0.2988	24.8658	11	0.8035	194.6825
37	0.2716	22.7482	12	0.7844	187.0684
38	0.2537	20.6306	13	0.7600	179.4543
39	0.2348	18.5130	14	0.7400	171.8402
40	0.2292	16.3954	15	0.7245	164.2261
41	0.2137	14.2778	16	0.7042	156.6120
42	0.2050	12.1602	17	0.6865	148.9979
43	0.2033	10.0426	18	0.6600	141.3838
44	0.1980	8.4311	19	0.6438	133.7697
45	0.1960	6.8196	20	0.6265	126.1556
46	0.1940	5.2081	21	0.6029	118.5415
47	0.1920	3.5966	22	0.5859	110.9274
48	0.1910	1.9851	23	0.5614	103.3133
49	**0.1905**	**0.8736**	24	0.5400	95.6992
50	**0.1901**	**0.2104**	25	0.5211	88.0851

في الحقبة الأولى، كان النموذج في بدايات تعلمه، حيث كانت قيمة معدل الخطأ في الكلمات (WER) مرتفعة جدًا وبلغت 0.9999، مما يشير إلى أن النموذج غير قادر بعد على التعرف بدقة على الكلام. ولكن مع استمرار عملية التدريب، بدأ الأداء يتحسن بشكل ملحوظ، حيث انخفض WER إلى ٠,٩٨٠٠ في الحقبة الثانية، مما يدل على أن النموذج بدأ في التعلّم من البيانات بشكل فعّال .مع مرور كل حقبة، استمر معدل الخطأ في الكلمات في الانخفاض بشكل تدريجي، حتى بلغ في الحقبة الأخيرة ٠,١٩٠١، أي ما يعادل ١٩% ، وهو معدل منخفض نسبياً يشير إلى قدرة النموذج العالية على التعرف على الكلمات بدقة مقبولة جداً. هذا الانخفاض يعكس فعالية تقليل عدد السمات في تسريع وتثبيت عملية التعلّم . أما بالنسبة لقيمة الخسارة(Validation Loss) ، فقد انخفضت تدريجياً من ٢٧٠,٨٢ في الحقبة الأولى إلى ٠,٢١٠٤ في الحقبة الأخيرة، وهو ما يشير إلى أن النموذج لا يعاني من الإفراط في التعلّم(Overfitting) ، بل أصبح قادراً على التعميم بشكل جيد عند التعامل مع بيانات لم يسبق له رؤيتها من قبل.

بشكل عام، يظهر أن أداء النموذج تحسّن بشكل كبير بمرور الوقت، ويؤكد أن استراتيجية تقليل عدد السمات إلى 53 سمة ساعدت على تسريع التدريب وتحسين جودة نتائج التعرّف الصوتي مع الحفاظ على مستوى جيد من الدقة.

٨- الاستنتاجات والتوصيات

في هذا البحث، تم تطوير نموذج محسن للتعرف التلقائي على الكلام باستخدام تقنيات التعلم العميق مع تقليل عدد السمات المستخرجة إلى ٥٣ سمة فقط. ومن خلال عملية التدريب والتقييم على مجموعة بيانات صوتية، توصلنا إلى النتائج التالية:

- تحقيق أداء جيد للنموذج: بلغ معدل خطأ الكلمات (WER) في نهاية التدريب ١٩,٠١% حيث كلما كان WERأقل، دلّ ذلك على أداء أقوى وأكثر دقة للنظام وهو يعتبر جيد مقارنة بالدراسات المرجعية الأخرى حيث بلغ معدل الخطأ بالكلمات 27 % [17]، مما يدل على أن النموذج استطاع التعرف على الكلمات بدقة مقبولة، رغم تقليل عدد السمات المدخلة. ويُظهر هذا أن النموذج قادر على تعميم الأنماط الصوتية والتعرف على الكلمات المنطوقة بكفاءة.
- فعالية تقليل عدد السمات: أثبت تقليل عدد السمات إلى ٥٣ سمة فقط أنه لا يضر بشكل كبير بدقة النموذج، بل ساعد في تقليل زمن التدريب واستهلاك الموارد الحاسوبية، مع المحافظة على جودة الأداء.
- تحسين واضح في :Val–loss تم الوصول إلى قيمة خسارة منخفضة بلغت ٠,٢١، مما يدل على قدرة النموذج على التعميم وعدم الوقوع في مشكلة التجهيز الزائد(Overfitting)
- أثر إيجابي للمعالجة الأولية: ساهمت مراحل المعالجة الأولية وزيادة البيانات في تعزيز أداء النموذج وتحسين قدرته على التعامل مع بيئات صوتية مختلفة.

التوصيات:

- يُنصح بالاستمرار في تطوير النموذج باستخدام هياكل أعمق أو هجينة تجمع بين نماذج التتابع مثل RNN و Transformer أو الاستفادة من تقنيات مثلAttention ويمكن استخدام تقنيات اختيار السمات (FeatureSelection)التكيفية لتحسين الأداء.

المراجع

[1] Reitmaier, T., Wallington, E., Kalarikalayil Raju, D., Klejch, O., Pearson, J., Jones, M., ... & Robinson, S. (2022, April). Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In Proceedings of the 2022 CHI conference on human factors in computing systems (pp. 1-17).
[2] Mao, C., & Liu, S. (2024). A Study on Speech Recognition by a Neural Network Based on English Speech Feature Parameters. Journal of Advanced Computational Intelligence and Intelligent Informatics, 28(3), 679-684.
[3] Nadeu, C., Macho, D., & Hernando, J. (2001). Time and frequency filtering of filter-bank energies for robust HMM speech recognition. Speech Communication, 34(1-2), 93-114.
[4] Jin, Z., Xie, X., Wang, T., Geng, M., Deng, J., Li, G., ... & Liu, X. (2024, April). Towards automatic data augmentation for disordered speech recognition. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10626-10630). IEEE.
[5] Liu, M., & Wei, Y. (2022). An improvement to conformer-based model for high-accuracy speech feature extraction and learning. Entropy, 24(7), 866.
[6] Nghia, H. N. H., Duy, N. T. H., Huy, N. G., Due, N. M. M., Luan, L. D., Hao, D. D., ... & Hung, V. T. (2022, October). Improving automatic speech recognition for low-resource language by data augmentation. In 2022 9th NAFOSTED Conference on Information and Computer Science (NICS) (pp. 159-164). IEEE.
[7] Geng, J., Jia, D., He, Z., Wu, N., & Li, Z. (2024). Enhanced Conformer-Based Speech Recognition via Model Fusion and Adaptive Decoding with Dynamic Rescoring. Applied Sciences, 14(24), 11583.
[8] Soni, M., Panda, A., & Kopparapu, S. K. (2024, December). Generalized SpecAugment: Robust Online Augmentation Technique for End-to-End Automatic Speech Recognition. In 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1-5). IEEE.
[9] Yang, Y., Wang, P., & Wang, D. (2022). A conformer based acoustic model for robust automatic speech recognition. arXiv preprint arXiv:2203.00725.
[10] https://keithito.com/LJ-Speech-Dataset/
[11] https://www.kaggle.com/datasets/ejlok1/cremad
[12] https://www.openslr.org/12
[13] Labied, M., Belangour, A., Banane, M., & Erraissi, A. (2022, March). An overview of automatic speech recognition preprocessing techniques. In 2022 international conference on decision aid sciences and applications (DASA) (pp. 804-809). IEEE.
[14] Zhou, X., Zhang, Y., Wang, Y., Tian, J., & Xu, S. (2024). Pyramid Feature Attention Network for Speech Resampling Detection. Applied Sciences, 14(11), 4803.
[15] Haoge, Z. H. A. N. G., Yubin, S. H. A. O., Hua, L. O. N. G., Yi, P. E. N. G., & Dachun, Z. H. O. U. (2023). Language Recognition Based on Log Gammatone-Scale Filter Bank Energies Spectrograms. Journal of Beijing University of Posts and Telecommunications, 46(1), 38.

[16] von Neumann, T., Boeddeker, C., Kinoshita, K., Delcroix, M., & Haeb-Umbach, R. (2023, June). On word error rate definitions and their efficient computation for multi-speaker speech recognition systems. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
[17] Tantawi, I. K., Abushariah, M. A., & Hammo, B. H. (2021). A deep learning approach for automatic speech recognition of The Holy Qur'ān recitations. International Journal of Speech Technology, 24(4), 1017-1032.