مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (8) العدد(3) 2024
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (8) No. (3) 2024

التعرف التلقائي على الكلام بالاعتماد على تقنيات التعلم العميق

د. م. راغب طعمه*

م. احمد حكمات محمد**

(تاريخ الإيداع 2024/1/21 . قُبِل للنشر في 2024/3/11)

□ ملخّص □

في عالمنا الحديث، أصبحنا أكثر اعتماداً على التكنولوجيا في كل جانب من جوانب حياتنا. فنحن نستخدمها للتواصل مع الآخرين، والعمل، والتعلم، والترفيه. وأحد أكثر المجالات التي شهدت تطوراً سريعاً في السنوات الأخيرة هي تقنية التعرف التلقائي على الكلام Automatic Speech Recognition (ASR) حيث تسمح هذه التقنية بتحويل اللغة المنطوقة إلى نص. يعدّ التعلم العميق Deep learning فرع من فروع الذكاء الاصطناعي فهو يعتمد على استخدام الشبكات العصبية الاصطناعية لمعالجة البيانات. وقد أثبتت الشبكات العصبية الاصطناعية أنها فعالة جدا في حل مجموعة متنوعة من المهام، بما في ذلك مهام التعرف التلقائي على الكلام ASR . في هذا البحث تم الاعتماد على تقنيات التعلم العميق في تصميم نموذج قادر على التعرف التلقائي على الكلام ASR في البيئات المختلفة وتحويلها إلى نص من خلال تدريبه على قاعدة بيانات كبيرة. يناقش البحث البنية الهيكلية للنموذج المقترح من حيث تحديد العدد الأمثل للطبقات الخفية والعدد الأمثل للعصبونات الموجودة في كل طبقة، فالنموذج المقترح عبارة عن بنية قوية لمهام التعرف على الكلام فهو يجمع بين نقاط قوة الطبقات التلافيفية والمتكررة لاستخراج ومعالجة الميزات المحلية والزمنية من الإشارة الصوتية وأظهرت النتائج المنجزة في بيئة بايثون Python الأداء الأفضل للنموذج المقترح المرتكز على نموذج الشبكة العصبونية. حيث حصلنا على معدل الخطأ في الكلمات Word Error Rate (WER) في الحقبة الأخيرة وهو 16 %. يعد معدل الخطأ هذا منخفضاً جداً مقارنة مع الدراسات الأخرى ويشير إلى أن النموذج قادر على التعرف على الكلام بدقة شديدة.

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

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

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

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (8) العدد(3) 2024
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (8) No. (3) 2024

Automatic Speech Recognition Based on Deep Learning Techniques

Dr. Ragheb Toghmah *
Ahmad Hikmat Mohammad**

(Received 31/1/2024 . Accepted 11/3/2024)

□ ABSTRACT □

In our modern world, we have become more dependent on technology in every aspect of our lives. We use it to communicate with others, work, learn, and entertain. One of the areas that has witnessed rapid development in recent years is Automatic Speech Recognition (ASR), as this technology allows spoken language to be converted into text. Deep learning is a branch of artificial intelligence, as it relies on the use of artificial neural networks to process data. Artificial neural networks have proven to be highly effective in solving a variety of tasks, including automatic speech recognition (ASR). In this research, we relied on deep learning techniques to design a model capable of automatic speech recognition (ASR) in different environments and converting it to text by training it on a large database. The research discusses the structural structure of the proposed model in terms of determining the optimal number of hidden layers and the optimal number of neurons present in each layer. The proposed model is a powerful architecture for speech recognition tasks, as it combines the strengths of convolutional and recurrent layers to extract and process local features. And the temporality of the audio signal. The results performed in the Python environment showed the best performance of the proposed model based on the neural network model. We obtained a Word Error Rate (WER) in the last era of 16%. This error rate is very low compared to other studies and indicates that the model is capable of recognizing speech very accurately.

Keywords: Deep Learning, Automatic Speech Recognition, Neural Networks, Word Error Rate, Feature Extraction, MFCC

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

١ – مقدمة:

يعدّ التعرف التلقائي على الكلام (ASR) أحد مجالات البحث العلمي ويهدف إلى تطوير تكنولوجيا قادرة على تحويل اللغة المنطوقة إلى نص مكتوب وقد شهد هذا المجال تطورات ملحوظة في السنوات الأخيرة، مما ساهم في تطبيقات تتراوح بين المساعدين الافتراضيين وخدمات النسخ إلى الأجهزة التي تعمل بالأوامر الصوتية وترجمة اللغات. يتضمن ASR في جوهره تطوير خوارزميات ونماذج تمكن الآلات من تحويل اللغة المنطوقة بدقة إلى نص، وبالتالي سد الفجوة بين التواصل البشري والأنظمة التكنولوجية. [1]

يتضمن البحث وراء التعرف التلقائي على الكلام نهجاً متعدد التخصصات ويستمد المعرفة والتقنيات من مجالات مختلفة مثل اللغويات ومعالجة الإشارات والتعلم الآلي وعلوم الكمبيوتر. كما يسعى العلماء والمهندسون جاهدين لمواجهة التحديات التي يمثلها تنوع وتعقيد الكلام، بما في ذلك لهجات المتحدث واللغات المختلفة وعدم طلاقة الكلام والضوضاء في الخلفية والتعرف على الكلام في سيناريوهات الوقت الحقيقي.

يعد التعلم العميق من أحدث التقنيات المستخدمة في مجال التعرف التلقائي على الكلام لما يبديه من دقة وسرعة وسهولة في عملية التعرف، أيضاً نماذج التعلم العميق قادرة على تعلم الأنماط المعقدة في إشارة الكلام التي لا يمكن التقاطها بسهولة بالطرق التقليدية. وذلك لأن نماذج التعلم العميق قادرة على تعلم التمثيلات الهرمية لإشارة الكلام، بدءاً من الميزات ذات المستوى المنخفض مثل تردد الإشارة وكثافتها إلى الميزات عالية المستوى مثل الصوتيات والكلمات التي تشكل الكلام. يمكن تدريب نماذج التعلم العميق على مجموعات كبيرة من بيانات الكلام وذلك لأن نماذج التعلم العميق قادرة على التعلم من كميات هائلة من البيانات وهو أمر ضروري لتحقيق دقة عالية في ASR.

٢ – الدراسات السابقة:

في عام 2021 قام الباحث Tantawi, I. K et al [2] بتصميم وتطوير وتقييم محرك ASR لتلاوات القرآن الكريم باستخدام نهج التعلم العميق في مجموعة أدوات KALDI وهي عبارة عن مجموعة أدوات مفتوحة المصدر للتعرف على الكلام ومعالجة الإشارات مكتوبة بلغة C++ حيث استخدم الباحث البنية المعمارية المؤلفة من الدخل وهو عباره عن كلام أي صوت حيث يتم استخراج الميزات باستخدام MFCC ومن ثم الحصول على شعاع الميزات بعدها حصل على النص من خلال القاموس بمقارنة النموذج الصوتي مع النموذج اللغوي وحصل على معدل خطأ بالكلمة WER بنسبة 27% [2]. كما قام الباحث Dhakal, P. et al في الدراسة [3] ببناء نظام التعرف التلقائي على الكلام في الزمن الحقيقي مستخدماً CNN مع بنية AlexNet لإستخراج الميزات ومن ثم تصنيفها باستخدام ثلاث مصنفات وهي SVM and RF and DNN وقام بعملية تصفية الضوضاء الغاوصية البيضاء المضافة Additive White Gaussian Noise (AWGN) التي تفسد إشارة الدخل الصوتية واستخدم طريقة التصفية التكيفية للمربعات الصغرى التكرارية Recursive Least Squares (RLS) لإلغاء الضوضاء. أما في عام 2020 استخدم الباحث Deshmukh, A. M. نموذجين لبناء نظام التعرف التلقائي على الكلام والمقارنة بينهما وهما نموذج الشبكات العصبية العميقة مع نماذج ماركوف المخفية DNN–HMM ونموذج الشبكات العصبية المتكررة المدربة على التصنيف الزمني الاتصالي RNN–CTC ووجد أن RNN أعطت نتائج أكثر دقة من ناحية معدل الخطأ بالكلمة بنسبة 26% وأسرع من ناحية التعرف على الكلام[4] .

اعتمدت الدراسة التجريبية[5] التي قام بها الباحث Ramadan, R et al على تقليل الضوضاء القائمة على الشبكة العصبية التلافيفية(CNN) في نظام التعرف على الكلام وكان الهدف الرئيسي إلغاء الضوضاء لتحسين إشارة الكلام والحصول على إشارة خالية من البقع بجودة أعلى، تم اعتماد الترشيح غير الخطي، حيث تمت تصفية 50٪ من الضوضاء. وجاء في الدراسة [6] أن أحد التحديات في أنظمة التعرف التلقائي على الكلام هو زيادة دقة وكفاءة نظام التعرف التلقائي على الكلام، وتتمثل إحدى طرق تحسين دقتها في تحسين النموذج الصوتي acoustic model (AM) وهو أحد مكونات نظام ASR المسؤول عن تحويل الخصائص الصوتية لإشارات الكلام ، واستخدام مزيج من احد أنواع شبكات العميقة Deep belief network (DBN) ، لاستخراج ميزات إشارات الكلام ، والذاكرة العميقة ثنائية الاتجاه طويلة المدى Deep Bidirectional Long Short-Term Memory (DBLSTM) مع طبقة الإخراج الخاصة بالتصنيف الزمني الاتصالي Connectionist Temporal Classification (CTC) لإنشاء النموذج الصوتي AM في مجموعة بيانات الكلام الفارسي.

٣- مشكلة البحث:

في سياق بعض اللغات، يعتبر التعرف التلقائي على الكلام تحدياً معقداً يتسم بالتنوع اللغوي الكبير وتباين التصوير الصوتي للكلمات والعبارات. ينبثق التحدي من التباين في اللهجات والنطق الذي يمتاز به هذا النوع من اللغات. يصبح تعرف النطق في تلك اللغات مهمة صعبة تمثل تحدياً أمام أنظمة التعرف التلقائي التقليدية. وتتزايد درجة التعقيد عند التعامل مع اللغات التي تفتقر إلى مجموعات بيانات صوتية وافرة ومتاحة لتدريب النماذج التعلم العميق.

هذه المشكلة تتجلى في تحقيق التواصل الفعال باستخدام الكلام في هذه اللغات، حيث تصبح العوامل المتعددة كاللهجات المتنوعة والنطق المتغير وعدم توفر البيانات الكافية عائقاً أمام دقة نظم التعرف التلقائي. وعلى الرغم من تطورات مجال التعلم العميق، ما زالت هناك حاجة ملحة لاستكشاف الحلول الفعالة لهذه المشكلة. وأيضاً تواجه أنظمة التعرف التلقائي على الكلام تحديات كبيرة عند تطبيقها في مجال الرعاية الصحية. يُعَدُّ هذا المجال حساساً ومعقداً، حيث تكمن أهمية دقة التعرف على الكلام في تحسين جودة الرعاية الصحية.

٤- أهمية البحث وأهدافه:

يقدم البحث مساهمة جديدة في مجال رفع كفاءة أنظمة التعرف التلقائي على الكلام من حيث تصميم نموذج هجين معتمد على الشبكات العصبونية التلافيفية متعددة الطبقات وعلى الشبكات العصبية المتكررة وذلك بهدف التعرف بشكل دقيق على الكلام وتحويله إلى نص بدقة عالية , وتكمن أهمية هذا البحث في تحديد أفضل هيكلية للنموذج المقترح من حيث تحديد العدد الأمثل للطبقات الخفية والعدد الأمثل للعصبونات الخفية الموجودة فيه بتقييم متوسط مربع الخطأ الناتج بعد كل عملية تدريب للشبكة العصبونية المبني عليها الموديل .يهدف هذا البحث إلى تقديم تقنية جديدة ومبتكرة في مجال رفع كفاءة أنظمة التعرف التلقائي على الكلام من خلال تصميم نموذج هجين متقدم يستند إلى تجميع مزايا الشبكات العصبونية التلافيفية متعددة الطبقات والشبكات العصبية المتكررة بهدف تحقيق دقة عالية في تحويل الكلام إلى نص مما يسهم في تحسين جودة أنظمة التعرف التلقائي على الكلام وتعزيز تجربة المستخدم. بالتالي يُعَدُّ هذا البحث مهم جدا في تقدم مجال تعرف الكلام وتحسين أدائه في مجموعة متنوعة من التطبيقات.

٥- طرق البحث ومواده:

يعتمد البحث على نموذجين أساسيين هما الشبكات العصبية التلافيفية والشبكات العصبية المتكررة. كما تم الاعتماد أيضاً على خوارزمية التصنيف الزمني الاتصالي Connectionist Temporal Classification من أجل تعيين تسلسل الإدخال إلى تسلسل الإخراج وتم بعد ذلك استخدام الخسارة لتحديث معلمات النموذج من خلال الانتشار العكسي. كما تم تطبيق تقنية Data Augmentation التي تشمل تنفيذ تغييرات بسيطة على البيانات الأصلية بطرق متعددة مثل إضافة ضجيج وزيادة سرعة مقطع صوتي والعديد من العمليات الأخرى. وتم استخدام تقنية Mel–Frequency Cepstral Coefficients لاستخراج الميزات الصوتية والتي تستخدم على نطاق واسع في تطبيقات مثل التعرف التلقائي على الكلام. ولتقييم دقة أنظمة التعرف التلقائي على الكلام (ASR) تم استخدام معدل خطأ الكلمات (WER) هو مقياس شائع الاستخدام لتقييم دقة هذه الأنظمة وغيرها من مهام التسلسل إلى التسلسل (Sequence–to–sequence).

٦- مراحل البحث

6-1 مرحلة الحصول على البيانات:

أحد العوامل المحورية في تعزيز أداء هذه الأنظمة هو توافر بيانات تدريبية واسعة النطاق ومتنوعة. في حين تساهم مجموعات البيانات الكبيرة بفروق لغوية قيّمة وتكاملها يمكن أن يؤدي إلى نموذج ASR أكثر قوة وقابلية للتعميم. يؤدي دمج مجموعات البيانات المتعددة إلى إثراء تعرض النموذج لأنماط الكلام واللهجات وضوضاء الخلفية المختلفة، مما يعزز قدرته على التكيف مع سيناريوهات العالم الحقيقي أيضاً يضمن دمج مصادر متعددة من البيانات تغطية أوسع للتنوع المعجمي واللغوي مما يؤدي إلى تحسين دقة التعرف عبر سياقات مختلفة. وفي هذا البحث تم الاعتماد على ثلاث مجموعات بيانات من أجل تنوع أنماط الكلام وهي:

6-1-1 The LJ Speech Dataset :

مجموعة بيانات كلامية ذات ملكية عامة تتكون تقريبا من 13 ألف مقطع صوتي قصير لمتحدث واحد يقرأ مقاطع من 7 كتب غير خيالية ويتم توفير النسخ لكل مقطع. تختلف مدة المقاطع من 1 إلى 10 ثوانٍ ويبلغ إجمالي طولها حوالي 24 ساعة، ويبلغ حجمها 2.6 GB ، نُشرت النصوص بين عامي 1884 و1964 وهي متاحة للعامة. تم تسجيل الصوت في 2016-2017 بواسطة مشروع LibriVox وهو أيضاً للملكية العامة.[7]

6-1-2 CREMA-Dataset :

هي عبارة عن مجموعة بيانات مكونة من 7442 مقطعاً أصلياً من 91 ممثلاً. كانت هذه المقاطع لـ 48 ممثلاً و43 ممثلة تتراوح أعمارهم بين 20 و74 عاماً ينتمون إلى مجموعة متنوعة من البلدان (أمريكا الأفريقية، الآسيوية، وغيرها) ويبلغ حجمها 473MB.[8]

6-1-3 LibriSpeech ASR :

هي عبارة عن مجموعة مكونة من 100 ساعة تقريباً من الخطاب باللغة الإنكليزية بسرعة 16 كيلو هرتز، أعدّها Vassil Panayotov بمساعدة Daniel Povey. هذه البيانات مستمدة من الكتب الصوتية المقروءة، ويبلغ حجمها 6.3 GB وقد تم تقسيمها ومواءمتها بعناية.[9]

6-2 النموذج الصوتي (Acoustic Model (AM:

النموذج الصوتي هو أحد مكونات نظام ASR المسؤول عن تحويل الخصائص الصوتية لإشارات الكلام إلى سلسلة من الصوتيات أو وحدات الكلمات الفرعية أو الوحدات اللغوية الأخرى. إنه في الأساس "المستمع" في عملية ASR. الغرض الأساسي من النموذج الصوتي هو التقاط العلاقة بين السمات الصوتية للغة المنطوقة والوحدات اللغوية التي تمثل تلك الأصوات. يأخذ AM سلسلة من الميزات الصوتية المستخرجة من الإشارة الصوتية (معاملات التردد الرأسي أو MFCCs) كمدخل ويخرج الاحتمالات أو الدرجات لمختلف الوحدات اللغوية، والتي يمكن أن تشمل الصوتيات، ثنائيات، ثلاثيات، أو حتى وحدات الكلمات الفرعية. تساعد هذه الاحتمالات النظام على تحديد التسلسل الأكثر احتمالاً للوحدات اللغوية التي تتوافق مع صوت الإدخال.

6-3 النموذج اللغوي Language Model (LM) :

يعد نموذج اللغة عنصراً حاسماً في ASR الذي يساعد النظام على التنبؤ باحتمالية وجود تسلسلات مختلفة من الكلمات أو الوحدات اللغوية في لغة معينة. إنه بمثابة "تنبؤ" بالكلمات أو الوحدات التي من المرجح أن تأتي بعد تسلسل معين. تعتمد LMs على تقنيات نمذجة اللغة الإحصائية وتتعلم التوزيع الاحتمالي لتسلسلات الكلمات في لغة معينة.

يأخذ LM النص كمدخل ويعين الاحتمالات لتسلسلات كلمات مختلفة. فهو يساعد نظام ASR على إزالة الغموض بين الكلمات أو الوحدات التي لها تمثيلات صوتية متشابهة ولكنها تختلف من حيث سياقها اللغوي. على سبيل المثال، عند مواجهة الميزات الصوتية لـ "two" و"too"، يمكن لـ LM مساعدة نظام ASR في تحديد أيهما أكثر احتمالاً بناءً على سياق الكلمات المحيطة.

يوضح الشكل (1) النموذج اللغوي والنموذج الصوتي حيث يتكون النموذج اللغوي من مجموعة من الكلمات مثلا five ويتكون النموذج الصوتي من اللفظ الذي يلفظه الأنسان لهذه الكلمات أي f – ay – v حيث تكون الية العمل عند إيجاد تسلسل الكلمات في الخرج نقوم بإيجاد النص المقابل لهذا التسلسل (اللفظ) :

الشكل (1):النموذج اللغوي والنموذج الصوتي

يعمل كلا النموذجين جنباً إلى جنب لتحسين دقة ASR من خلال الجمع بين المعلومات الصوتية واللغوية، أيضاً تم تحميل كل منهما من مختبر تكنولوجيا وأبحاث النطق Speech Technology and Research(STAR) [10] .

6-4 الطريقة المقترحة

تم في هذا البحث بناء نموذج قادر على التعرف التلقائي على الكلام من خلال تحويل الكلام إلى نص باستخدام مجموعة من الخطوات أهمها زيادة البيانات Data Augmentation والتي تعتبر خطوة مهمة في عملية تعرّض النموذج لأنماط الكلام واللّهجات وضوضاء الخلفية مما يعزز قدرته على التكيف مع سيناريوهات العالم الحقيقي والذي لم يتم التطرق لها في الدراسات السابقة، وأيضاً استخراج الميزات Feature Extraction ذات الصلة من كل إطار والعمل على تحسين هذا الميزات باستخدام طرق التحسين مثل Feature Normalization لضمان توافق مجموعات البيانات المستخدمة وأيضاً لم يتم التطرق له في الدراسات السابقة . وفيما يلي الخطوات المتبعة لبناء النموذج:

1- المعالجة الأولية للمقاطع الصوتية Pre-Processing

2- زيادة البيانات Data Augmentation

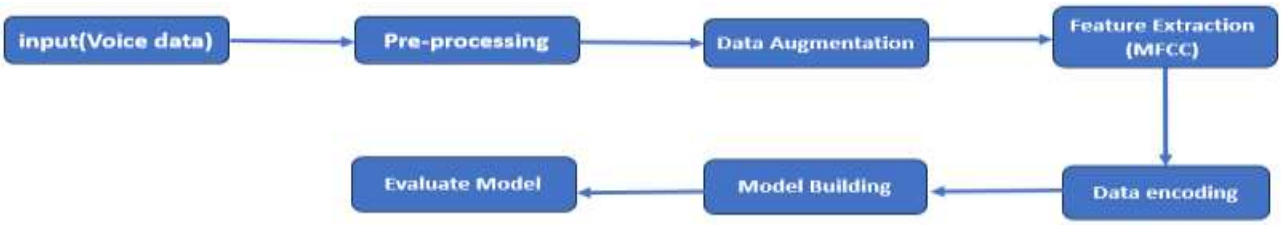
3- استخراج الميزات باستخدام طريقة معاملات ميل التردد الرأسي MFCC

4- ترميز البيانات (الصوت والنص)

5- بناء النموذج

6- تقييم أداء النموذج

ويوضح الشكل (2) هذه الخطوات حيث يتم معالجة البيانات قبل زيادتها، وذلك لتجنب إعادة معالجة البيانات بعد الزيادة ، حيث قد تكون قد تم معالجتها مسبقاً. سنوضح أيضاً كيفية عملية زيادة البيانات بشكل كامل في الفقرات التالية.



الشكل (2): الخطوات المتبعة لبناء النموذج المقترح

6-4-1: المعالجة الأولية للمقاطع الصوتية Pre-Processing:

تعد المعالجة المسبقة للصوت خطوة أساسية وحاسمة في التعامل مع البيانات الصوتية لمجموعة واسعة من التطبيقات. يتضمن سلسلة من العمليات والتحويلات المطبقة على الإشارات الصوتية الأولية لجعلها مناسبة للتحليل واستخراج الميزات والمعالجة الإضافية وفيما يلي قائمة بالخطوات التي تم إجرائها في المعالجة المسبقة للصوت:

1- أخذ العينات Sampling: يتم تحويل الإشارات الصوتية المستمرة إلى عينات رقمية منفصلة باستخدام معدل أخذ العينات المحدد (44.1 khz).

2- التطبيع Normalization: يتم ضبط سعة الصوت على نطاق قياسي (-1 إلى 1) يُعتبر هذا النطاق مألوفا ومستخدما على نطاق واسع في العديد من تطبيقات المعالجة الصوتية لضمان تحقيق مستويات صوت متسقة.

3- تحويل الملفات الصوتية Audio File Conversion: تحويل كامل الملفات الصوتية إلى تنسيق واحد wav من أجل سهولة التعامل مع هذه الملفات.

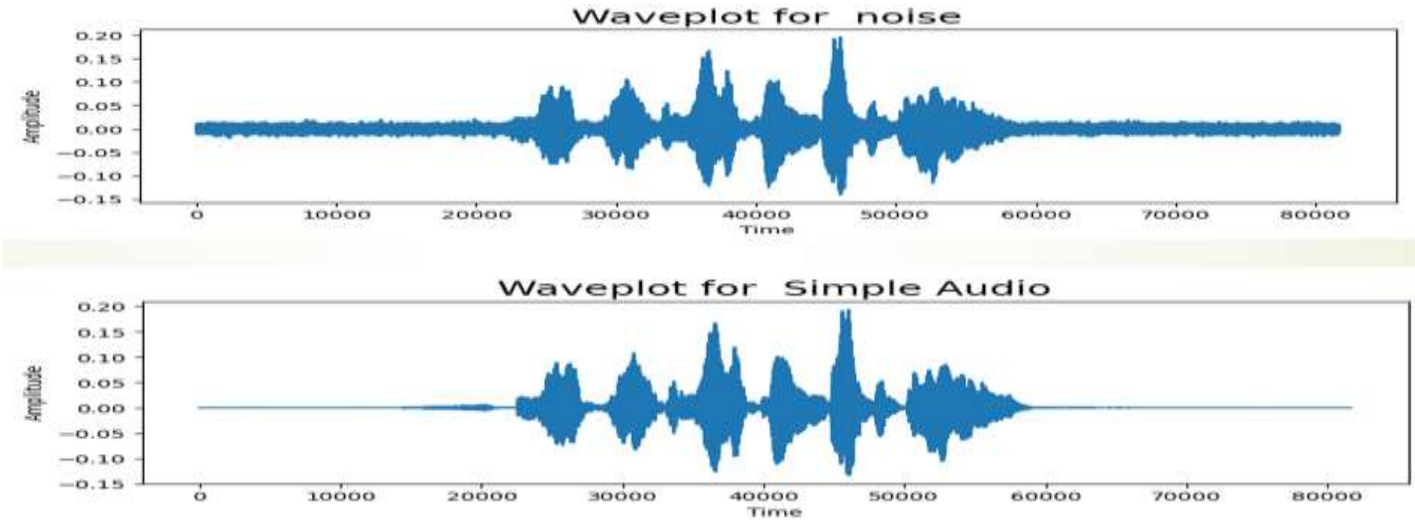
تعد المعالجة المسبقة للصوت خطوة حاسمة في مسار التحليل الصوتي الذي يقوم بإعداد البيانات الصوتية الأولية لمختلف التطبيقات، بما في ذلك التعرف على الكلام وتحليل الموسيقى وتصنيف الصوت والمزيد. من خلال معالجة مشكلات مثل الضوضاء والتطبيع واستخراج الميزات، تساعد المعالجة المسبقة على تحسين جودة وموثوقية مهام التحليل والنمذجة اللاحقة.

6-4-2 زيادة البيانات Data Augmentation:

تعتبر تقنيات زيادة البيانات خطوة مهمة جداً في نماذج التعرف التلقائي على الكلام لأنها تساعد على تحسين قوة نماذج التعلم الآلي والتعلم العميق من خلال إدخال الاختلافات والتنوع في بيانات التدريب. تتضمن هذه التقنيات تطبيق عمليات مختلفة على البيانات الصوتية الأصلية لإنشاء نسخ معززة منها. وفيما يلي شرح كل من العمليات المستخدمة في زيادة البيانات وتعزيزها:

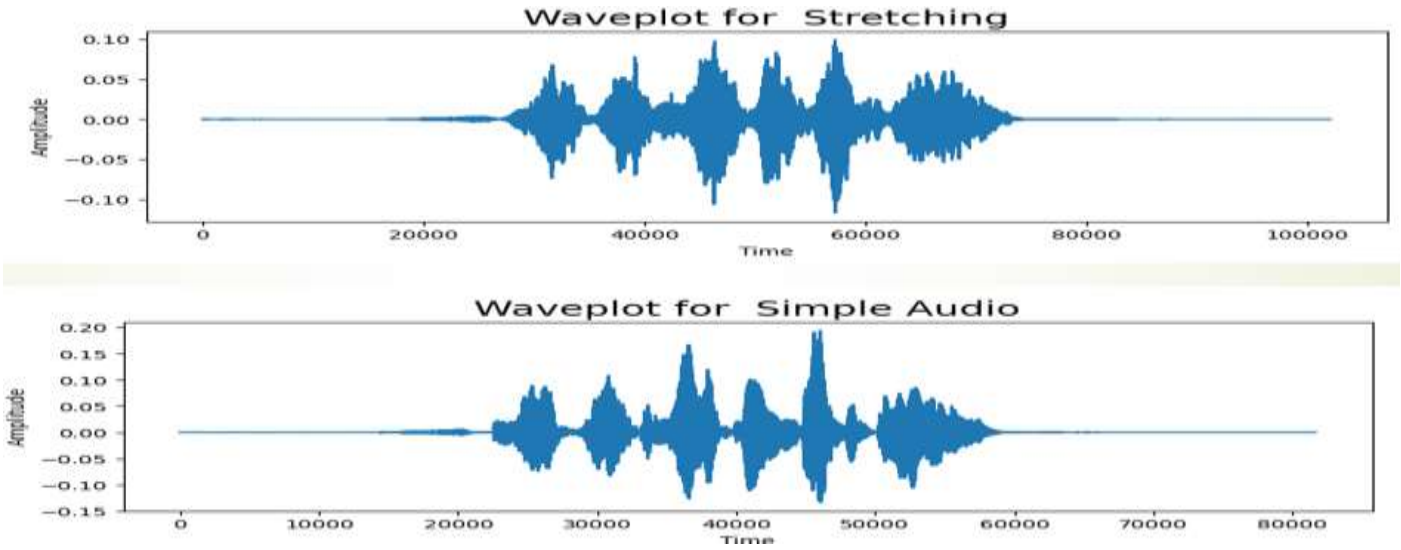
1- **إضافة ضوضاء Noise**: يتم إدخال ضوضاء عشوائية للإشارة الصوتية، وهي عبارة عن ضوضاء بيضاء، ويمكن إضافة أنواع أخرى من الضوضاء. حيث تساعد إضافة الضوضاء في أن يصبح النموذج أكثر قوة في مواجهة الضوضاء أو التداخل البيئي. بحيث يحاكي تحديات البيانات الصوتية في العالم الحقيقي مما يجعل النموذج أفضل في التمييز بين الإشارة والضوضاء، نلاحظ من الشكل (3) تغير شكل الإشارة الصوتية بعد إضافة الضوضاء. التغيرات في الشكل تشمل تغير في الطيف الصوتي، وتشويهات ترددية، مما يجعل الإشارة الناتجة تتباين عن الإشارة الأصلية وتصبح أكثر تعقيدا وتحديا للنموذج في التعرف عليها.

2- **التمدد Stretching**: يتضمن التمدد (أو التمدد الزمني) تغيير مدة الإشارة الصوتية مع الحفاظ على درجة



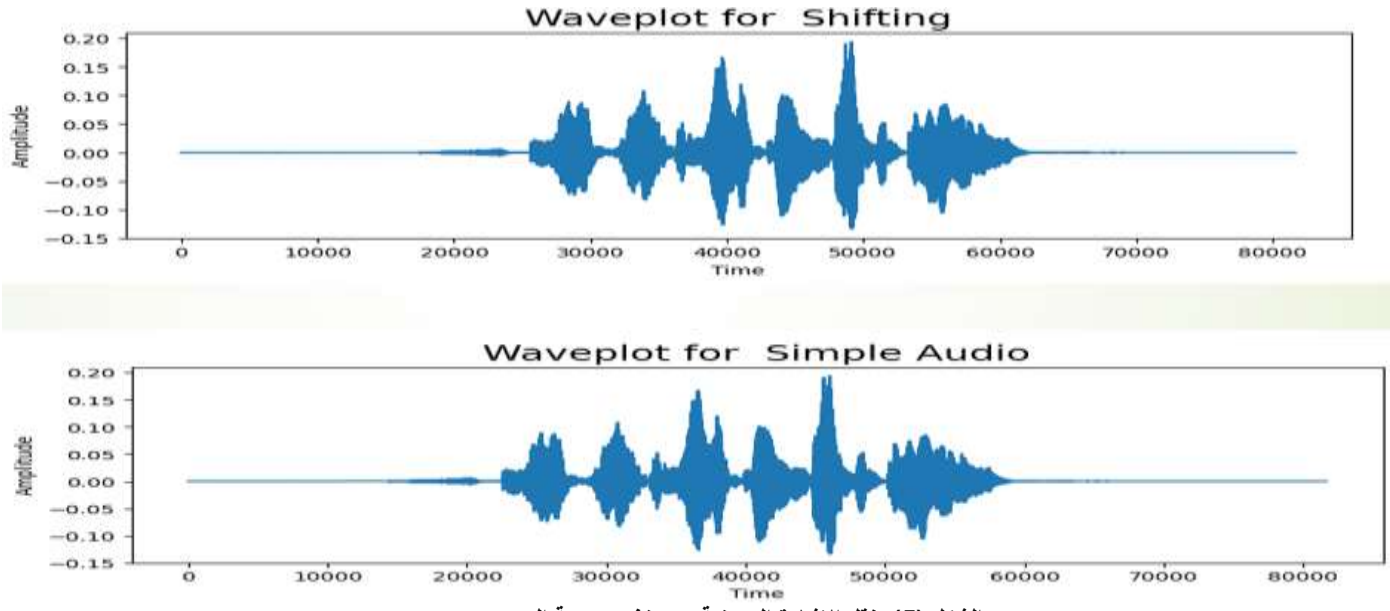
الشكل (3): تغير شكل الإشارة الصوتية بعد إضافة ضوضاء

الصوت كما هو موضح في الشكل (4). ويتم تحقيق ذلك عن طريق إعادة تشكيل الصوت بمعدل زمني مختلف. فالتمدد يحاكي الاختلافات في سرعة التحدث وهو أمر مفيد لمهام التعرف على الكلام كما يساعد النموذج على التكيف مع إيقاعات التحدث المختلفة. ويوضح الشكل (4) ان الإشاره الاصلية قد انتهت عند المدة الزمنيه 6000 وبعد اضافه تمدد أصبحت تقريبا تنتهي عند 7500



الشكل(4): تغيير مدة الإشارة الصوتية مع الحفاظ على درجة الصوت

3- الإزاحة Shifting: يتضمن نقل الإشارة الصوتية أي تحريكها على طول محور الوقت دون تغيير درجة الصوت كما هو موضح في الشكل (5). يتم تحقيق ذلك عن طريق إضافة أو إزالة الصمت في بداية الإشارة أو نهايتها حيث الإشارة الصوتية تبدأ عند 2200 وبعد الإزاحة أصبحت تبدأ عند 2700. يمكن أن يحاكي النقل نقاط بداية مختلفة

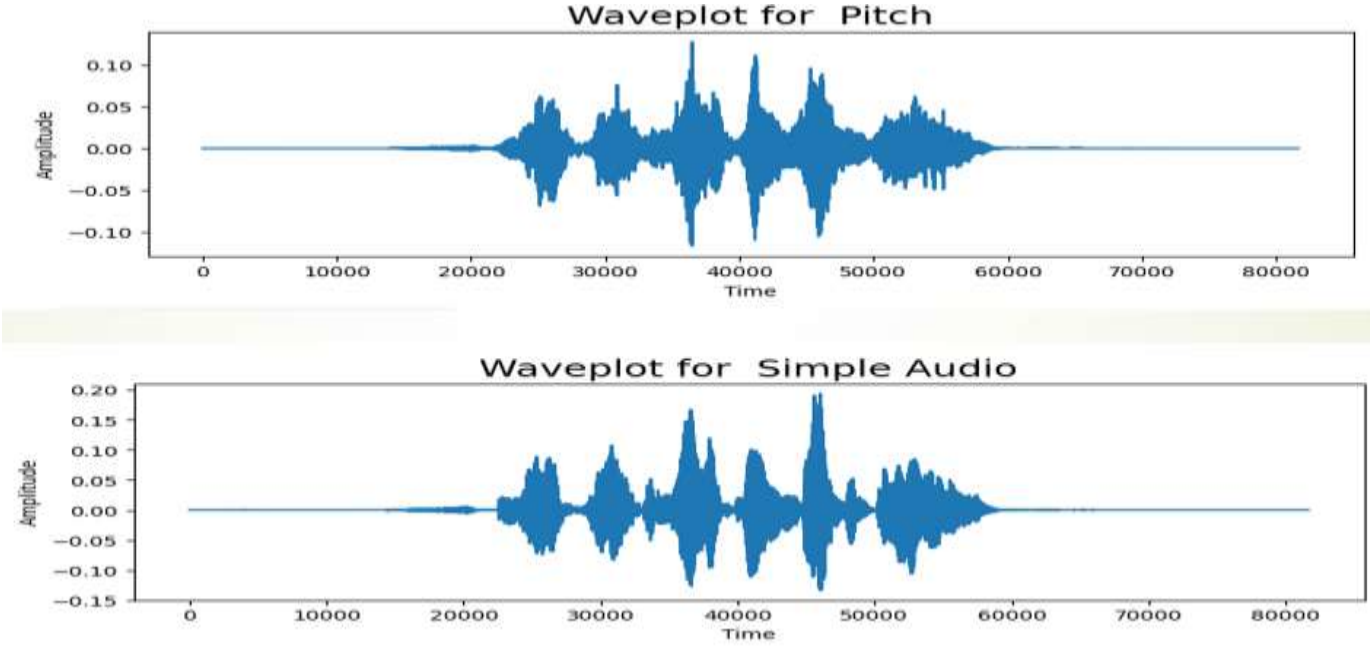


الشكل (5): نقل الإشارة الصوتية دون تغيير درجة الصوت

أو إزاحات زمنية في التسجيلات الصوتية. وهذا مفيد لمهام مثل تحليل المشهد الصوتي وقد يحدث نفس الصوت في أوقات مختلفة.

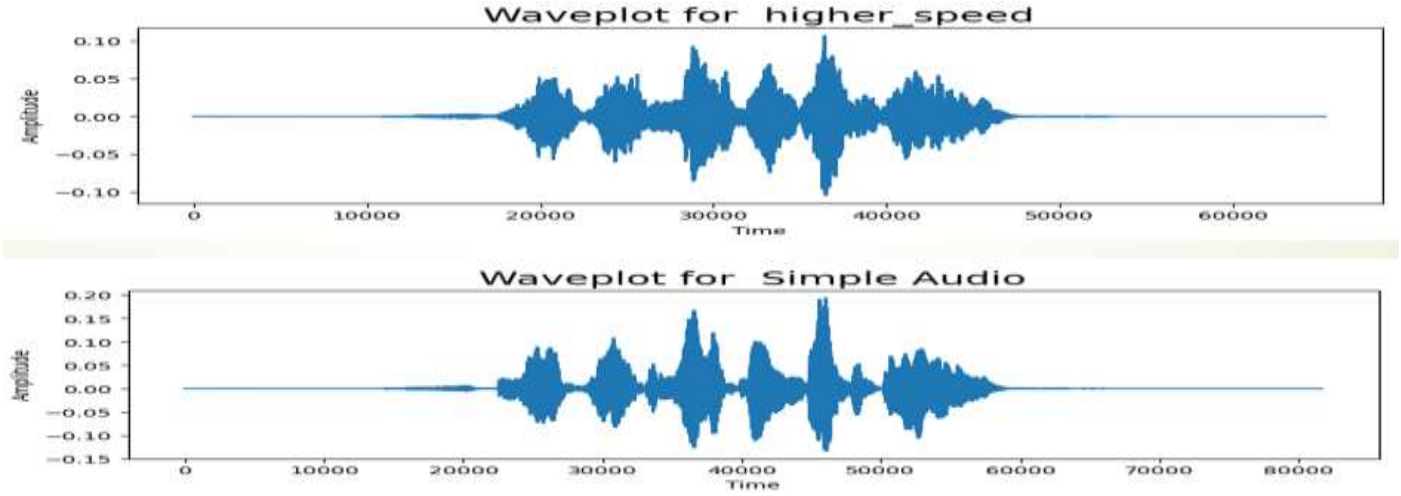
4- الاهتزاز Pitch: يتضمن تغيير الدرجة الصوتية أي تغيير التردد الأساسي للصوت بشكل بسيط جداً مع الحفاظ على مدته كما هو موضح في الشكل (6). ويتم تحقيق ذلك عن طريق قياس محور الوقت بشكل مختلف لمكونات

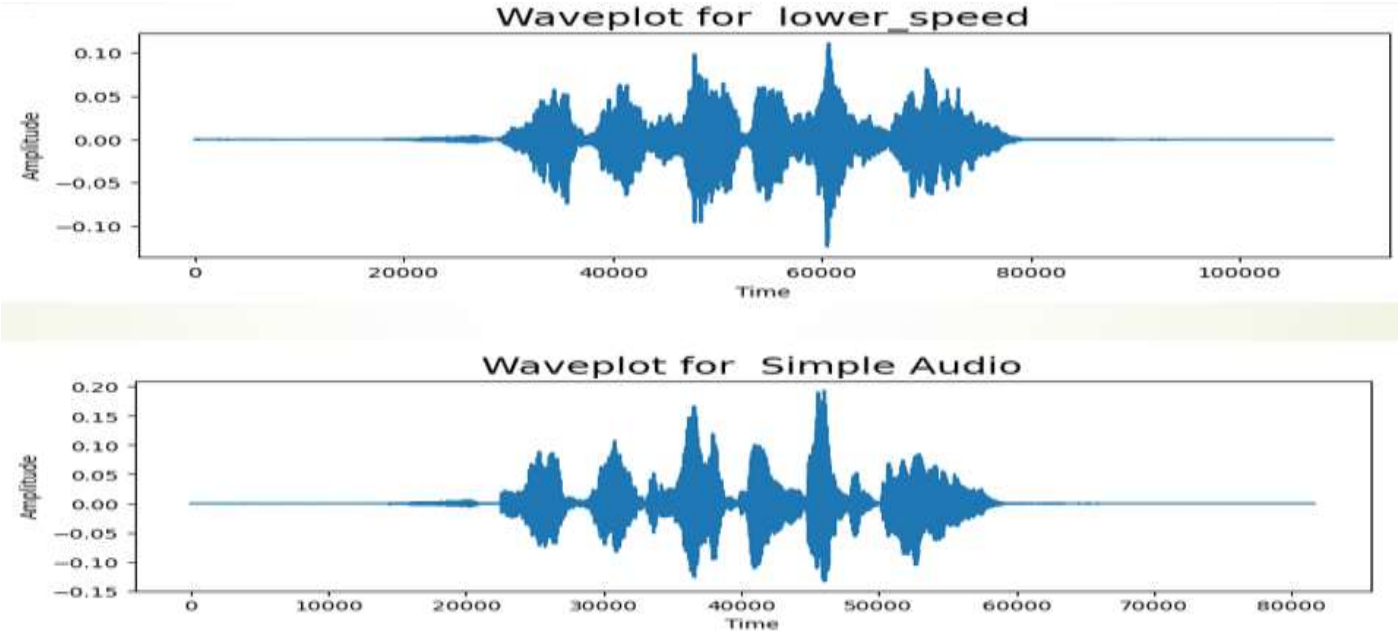
التردد المختلفة، تغيير درجة الصوت يمكن أن يحاكي الاختلافات في درجة الصوت أو النوتات الموسيقية. وهذا يكون مفيد لمهام مثل التعرف على المتحدث حيث قد يتحدث الأفراد بنبرة مختلفة قليلاً.



الشكل (6): تغير الدرجة الصوتية مع الحفاظ على مدته

5- سرعة أعلى وسرعة أقل Higher Speed or Lower Speed: يتم ضبط سرعة تشغيل الإشارة الصوتية حيث السرعة العالية تجعل تشغيل الصوت أسرع مع تقليل الوقت كما في الشكل (7)، بينما السرعة المنخفضة تجعل تشغيله أبطأ مع زيادة الوقت كما في الشكل (8). وهذا يمكن أن يحاكي اختلاف السرعة سيناريوهات العالم الحقيقي حيث يتم تسجيل الصوت أو تشغيله بسرعات مختلفة. وهذا مفيد جداً لنماذج التدريب للتعامل مع الاختلافات في أجهزة التشغيل أو ظروف التسجيل أو الأشخاص التي تتكلم بسرعة أو ببطيء.



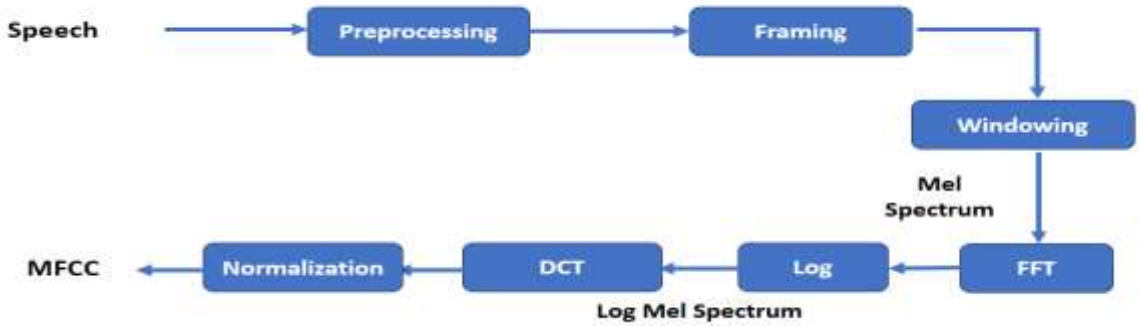


الهدف الأساسي من زيادة البيانات في المعالجة المسبقة للصوت هو تعرّض نموذج التعلم العميق لمجموعة واسعة من المدخلات المحتملة، مما يجعله أكثر مرونة في مواجهة الاختلافات والضوضاء التي قد يواجهها أثناء التعرف. تسمح هذه التقنيات للنموذج بالتعميم بشكل أفضل والأداء الجيد على البيانات غير المرئية، مما يؤدي في النهاية إلى تحسين دقة النموذج وفعاليته لمهام التعرف على الكلام.

3-4-6 : استخراج الميزات Feature Extraction : يمكن تعريف استخراج الميزات بأنها عملية تحويل البيانات الأولية إلى تمثيل أكثر إحكاماً وذات معنى يمكن استخدامه لمزيد من التحليل. في سياق التعرف التلقائي على الصوت يتم استخدام استخراج الميزات لتحويل الإشارة الصوتية الخام إلى مجموعة من الميزات التي يمكن استخدامها لتحديد المتحدث أو الكلمات المنطوقة.

إحدى تقنيات استخراج الميزات الأكثر شيوعاً المستخدمة في التعرف التلقائي على الكلام هي معاملات ميل التردد الرأسي Mel Frequency Cepstral Coefficients (MFCC) [13] وهي عبارة عن مجموعة من المعاملات التي تمثل الخصائص الطيفية للإشارة الصوتية. يتم حسابها عن طريق أخذ تحويل فوريه للإشارة الصوتية أولاً، ثم تطبيق بنك مرشح Mel على طيف الطاقة الناتج، وأخيراً أخذ لوغاريتم خرج بنك المرشح.

حيث أن بنك مرشح Mel عبارة عن مجموعة من المرشحات المثلثة المتباعدة لوغاريتمياً في مقياس Mel وهو مقياس قائم على الإدراك الحسي ويأخذ في الاعتبار الطريقة التي يسمع بها البشر الصوت. وهذا يجعل MFCCs أكثر قوة في مواجهة الاختلافات في درجة صوت السماعة وجهارة الصوت. يوضح الشكل (9) خطوات استخراج الميزات باستخدام MFCC:



الشكل (9): خطوات استخراج الميزات باستخدام MFCC

pre-processing: تعد المعالجة المسبقة للصوت خطوة حاسمة في إعداد البيانات الصوتية وتتم بمجموعة من الخطوات الشائعة في المعالجة المسبقة للصوت مثل أخذ العينات Sampling بحيث يتم تحويل الإشارات الصوتية المستمرة إلى عينات رقمية منفصلة باستخدام معدل أخذ العينات المحدد (44.1 كيلو هرتز) ،أيضاً التطبيع Normalization بحيث يتم ضبط سعة الصوت على نطاق قياسي (−1 إلى 1) لضمان مستويات صوت متسقة.

تجزئة الإطار Frame Segmentation: يتم تقسيم الإشارة الصوتية إلى إطارات أو نوافذ أصغر (20−50 مللي ثانية) للتحليل.

النوافذ Windowing: تطبيق وظيفة النافذة (هامينج Hamming) على كل إطار لتقليل التسرب الطيفي عند حواف الإطار.

تحويل فورييه السريع Fast Fourier Transform (FFT): حساب تحويل فورييه السريع لكل إشارة للحصول على تمثيل مجال التردد للإشارة.

مقياس ميل Mel Filter: مقياس ميل هو مقياس إدراكي للطبقات التي تقارب استجابة الأذن البشرية لترددات مختلفة.

اللوغاريتم Logarithm: أخذ لوغاريتم النتائج من مرشح Mel وهذا يساعد في محاكاة الإدراك البشري لجهارة الصوت.

تحويل جيب التمام المنفصل Discrete Cosine Transform (DCT) : يتم تطبيق DCT على مخرجات تصفية السجل حيث يقوم DCT بإلغاء ربط القيم ويلتقط أهم الميزات.

التطبيع Normalization: يتم تطبيع متجهات الميزات لجعلها أكثر قوة في مواجهة الاختلافات في ظروف الإدخال.

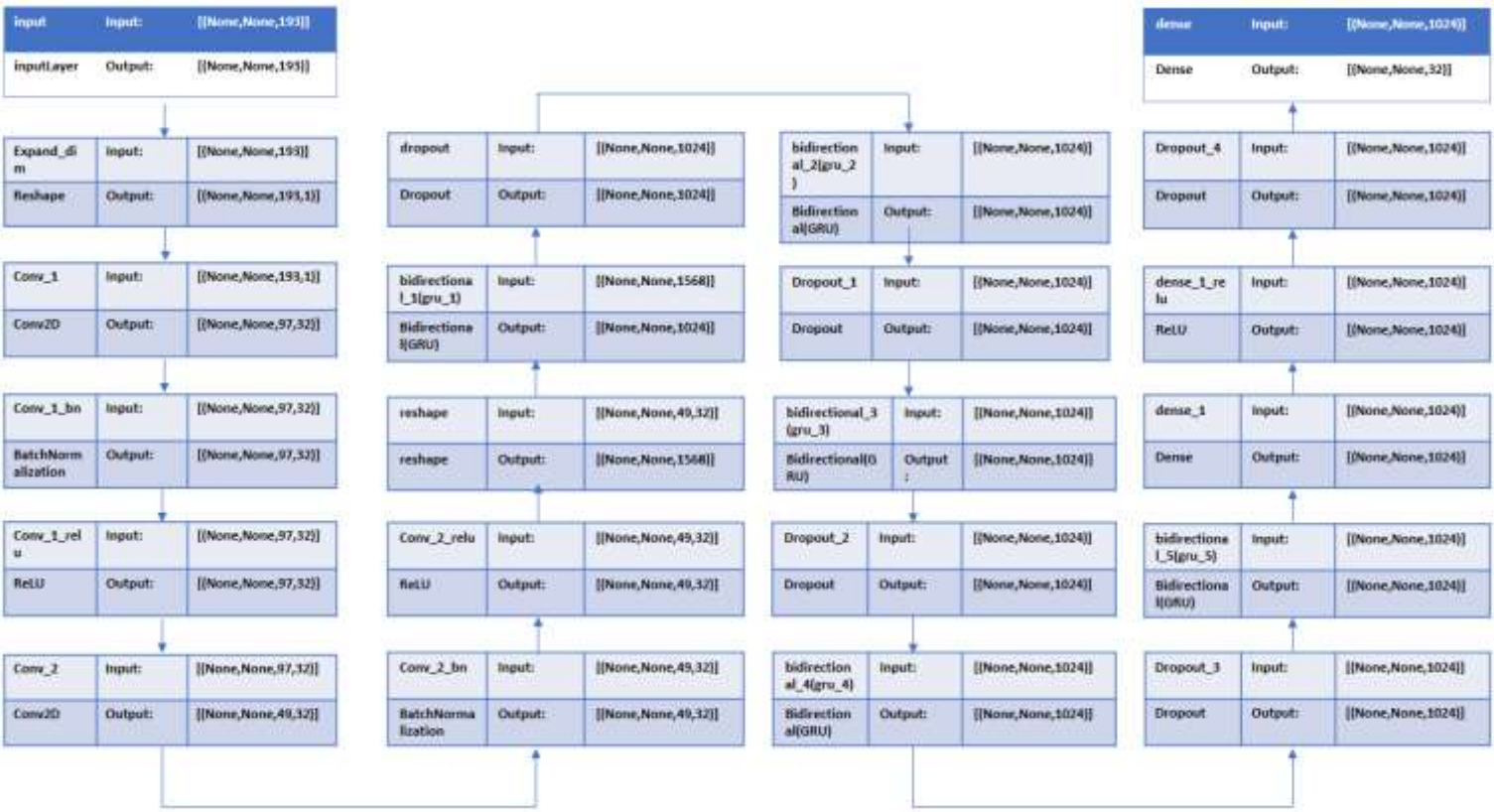
MFCC feature: تتوفر لدينا سلسلة من ناقلات ميزات MFCC التي يمكن استخدامها لمختلف مهام معالجة الإشارات الصوتية مثل التعرف على الكلام.

6-4-4 ترميز البيانات (الصوت والنص):

في أنظمة التعرف التلقائي على الكلام (ASR)، يشير ترميز البيانات إلى تمثيل الإشارات الصوتية (الميزات) والمعلومات النصية بتنسيق مناسب للتدريب.

- **ترميز الصوت Audio Encoding:** استخرج MFCCs من الإشارة الصوتية الأولية كما هو موضح سابقاً.
- **ترميز النص Text Encoding:** تعيين كل حرف في التسمية إلى تمثيل رقمي باستخدام دالة تسمى char_to_num.

6-4-5 بناء الموديل:



يحدد الشكل (10) نموذج الشبكة العصبية المتكررة التلافيفية convolutional recurrent neural network (CRNN) لمهام التعرف على الكلام، فالنموذج المقدم عبارة عن بنية قوية لمهام التعرف على الكلام فهو يجمع بين نقاط قوة الطبقات التلافيفية والمتكررة لاستخراج ومعالجة الميزات المحلية والزمنية من الإشارة الصوتية. تمكن الطبقات المتكررة ثنائية الاتجاه النموذج من التقاط السياق الأمامي والخلفي وهو أمر بالغ الأهمية للنسخ الدقيق للكلام. ويتكون النموذج من ثلاثة مكونات رئيسية:

1. **الطبقات التلافيفية:** تستخرج هذه الطبقات الميزات من مخطط طيف الإدخال، وهو تمثيل للإشارة الصوتية.
2. **الطبقات المتكررة:** تعالج هذه الطبقات الميزات المستخرجة وتلتقط المعلومات التسلسلية من إشارة الكلام.
3. **الطبقات الكثيفة:** تقوم هذه الطبقات بتحويل مخرجات الطبقات المتكررة إلى توزيع احتمالي على الأحرف المحتملة في المفردات.

6-4-5-1 بنية النموذج Model Architecture:

- طبقة الدخل: مدخل النموذج هو مخطط طيفي يمثل طيف تردد الإشارة الصوتية.
- الطبقة التلافيفية 1: تطبق هذه الطبقة عملية تلافيفية ثنائية الأبعاد مع 32 مرشحاً، وحجم نواة 41*11، وخطوة 2*2 . تم ضبط الحشو على نفسه للحفاظ على الأبعاد المكانية. يتم تطبيق تطبيع الدفعة وتنشيط ReLU بعد الالتفاف.
- الطبقة التلافيفية 2: هذه الطبقة تشبه الطبقة التلافيفية الأولى، ولكن بحجم نواة أصغر يبلغ 21*11 وخطوة 2*1. وتهدف هذه الطبقة إلى استخراج ميزات أكثر دقة من الإشارة الصوتية.
- الطبقات المتكررة: يعالج هذا الجزء من النموذج المعلومات الزمنية في إشارة الكلام. وهو يتألف من عدة وحدات متكررة. في هذه الحالة فإن الوحدة المتكررة هي وحدة متكررة ببوابات gated recurrent unit (GRU)، وهي خيار شائع لمهام التعرف على الكلام نظراً لقدرتها على التقاط التبعيات طويلة المدى.
- RNN ثنائي الاتجاه: يتم تكوين كل وحدة متكررة للعمل في الوضع ثنائي الاتجاه، مما يعني أنها تعالج تسلسل الإدخال للأمام والخلف وهذا يسمح للنموذج بالتقاط السياق الأمامي والخلفي لإشارة الكلام.
- التسربDropout : يتم استخدام طبقات التسرب لتسوية النموذج ومنع الإفراط في التجهيز. ويؤدي التسرب إلى إزالة جزء من وحدات الإدخال بشكل عشوائي أثناء التدريب مما يجبر النموذج على تعلم ميزات أكثر قوة.
- الطبقة الكثيفة Dense Layer : تأخذ هذه الطبقة مخرجات الطبقات المتكررة وتحولها إلى 2D tensor.
- تنشيط ReLU: يتم تطبيق تنشيط ReLU على مخرجات الطبقة الكثيفة لإدخال عدم الخطية non-linearity في النموذج.
- التسربDropout : يتم استخدام طبقة التسرب الأخرى لزيادة تنظيم النموذج.
- طبقة الإخراج: الطبقة الأخيرة عبارة عن طبقة كثيفة تحتوي على وحدات1 + put_dim، حيث يمثلput_dim حجم المفردات (أي عدد الأحرف الفريدة). ويتم تطبيق تنشيط Softmax على طبقة الإخراج لإنشاء توزيع احتمالي على الأحرف المحتملة.

6-4-5-2 تجميع النموذج Model Compilation :

يتم تجميع النموذج باستخدام مُحسِّن Adam ووظيفة CTC loss. حيث يُعد CTC loss خياراً مناسباً لمهام التعرف على الكلام لأنه يفسر الغموض المتأصل في المحاذاة بين تسلسلات الأحرف وتسلسلات الوحدات الصوتية.

6-4-5-3 معدل خطأ الكلمات Word Error Rate (WER):

هناك العديد من المقاييس التي يمكن استخدامها لتقييم أداء أنظمة ASR. والمقياس الأكثر شيوعاً هو معدل خطأ الكلمات (WER) ويعرف بأنه النسبة المئوية للكلمات الموجودة في النص المرجعي والتي لم يتعرف عليها نظام ASR بشكل صحيح [11] . ويشير انخفاض WER إلى أداء أفضل. ويحسب باستخدام الصيغة التالية:

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

حيث:

S: هو عدد البدائل (الكلمات غير الصحيحة). D هو عدد المحذوفات (الكلمات المفقودة).

I هو عدد الإدخالات (الكلمات الإضافية). N هو العدد الإجمالي للكلمات في النص المرجعي.

لحساب WER، تتم مقارنة النص المرجعي والنص الذي أنشأه النظام كلمة بكلمة وأي اختلافات بين الكلمات تعتبر أخطاء. ويمكن تصنيف هذه الأخطاء إلى ثلاثة أنواع:

1) البدائل: يقوم النظام باستبدال الكلمة الصحيحة في النص المرجعي بكلمة غير صحيحة.
2) عمليات الإدراج: يضيف النظام كلمة غير صحيحة إلى النص المرجعي.
3) عمليات الحذف: يحذف النظام الكلمة الصحيحة من النص المرجعي.

يتم بعد ذلك جمع عدد الاستبدالات والإدراجات والحذف وتقسيمها على إجمالي عدد الكلمات في النص المرجعي وهذا يعطي WER معبراً عنه كنسبة مئوية، وكما ذكرنا سابقاً يشير انخفاض WER إلى أداء أفضل ويعني WER بنسبة 0% أن النظام قام بنسخ كل كلمة في النص المرجعي بشكل صحيح.

4-5-4-6 وظيفة الخسارة Loss Function:

في سياق التعرف التلقائي على الكلام (ASR)، تعد وظيفة الخسارة مكوناً حاسماً يستخدم أثناء تدريب النموذج. الهدف من تدريب نظام ASR هو تقليل هذه الخسارة مما يحدد الفرق بين المخرجات المتوقعة للنموذج والحقيقة الفعلية.

CTC Loss (Connectionist Temporal Classification):

تتعامل خسارة CTC، التي تحظى بشعبية خاصة في ASR، مع مشكلات التسلسل إلى التسلسل حيث لا تكون المحاذاة بين المدخلات والمخرجات معروفة مسبقاً. ويسمح للنموذج بتعلم المحاذاة أثناء التدريب. غالباً ما تكون المحاذاة بين إطارات صوت الإدخال ورموز الإخراج (الصوتيات أو الأحرف أو وحدات الكلمات الفرعية) متغيرة وتعتمد على السياق. تعالج CTC هذا التحدي من خلال النظر في جميع التوافقات الممكنة أثناء التدريب، يتم حساب CTC Loss كما هو مبين في المعادلة (2)[12]:

$$L(X,Y) = -\log P(Y|X) \qquad (2)$$

حيث: X: تسلسل إدخال معين.

Y: تسلسل الإخراج.

تعد CTC Loss أداة قوية في ASR مما يسمح للنماذج بتعلم التعيينات المعقدة بين التسلسلات ذات الطول المتغير. حيث أن مرونته وقدرته على التعامل مع المحاذاة المتغيرة تجعله مكوناً رئيسياً في تدريب أنظمة ASR الشاملة.

7- النتائج والمناقشة:

يوضح هذا الجزء من البحث نتائج عملية المعالجة المسبقة لسمات مجموعة البيانات الضخمة بالتفصيل ومن ثم نتائج عملية التعرف باستخدام موديل التعلم العميق حيث تم تنفيذ العمل باستخدام البيئة البرمجية python (v3.8.3) .

7-1 سمات مجموعة البيانات:

لدينا 59 سمة من البيانات الصوتية جدول (1) مما يوفر تمثيلاً تفصيلياً للخصائص الصوتية للإشارة الصوتية. تشمل هذه السمات كلاً من الواصفات الأساسية للمحتوى الطيفي للإشارة الصوتية، مثل MFCCs، والنقطة الوسطى الطيفية، والتدحرج الطيفي، بالإضافة إلى الميزات المشتقة من معالجة الإشارة الصوتية، مثل MFCCs المضافة للضوضاء، وMFCCs الممتدة، وMFCCs المتغيرة درجة الصوت. تتيح هذه المجموعة الشاملة من السمات تحليل التعرف على الصوت. تلتقط السمات المستخرجة جوهر جرس الإشارة الصوتية وإيقاعها مما يسمح بفهم دقيق لمحتواها وبنيتها العامة.

جدول (1) : سمات مجموعة البيانات

السعة الإجمالية للطيف	MFCC 1
المعاملات الرأسية ذات التردد الميلي التي تلتقط الخصائص الطيفية	MFCC 2➔ 13
طاقة الإشارة	MFCC 14
معاملات الترتيب الأعلى تلتقط التفاصيل الدقيقة	MFCC 15➔ 59

7-2 تقليل السمات:

يشير تقليل السمات في سياق التعلم العميق إلى عملية اختيار مجموعة فرعية من السمات الأكثر صلة وأهمية من المجموعة الأصلية، وذلك بهدف تحسين أداء النموذج، وتقليل التعقيد الحسابي. ويمكن للسمات الإضافية أن تزيد من زمن الحساب كما يمكنها التأثير على دقة التعرف وبالتالي عملية تقليل السمات عملية ضرورية، وتم الاعتماد في هذا البحث على علاقة رياضية تساعد في إيجاد السمات الغير مفيدة من خلال حساب التباين، تضم هذه المرحلة خطوة أساسية وهي:

7-2-1 حساب التباين:

التباين هو قياس لدرجة تشتت البيانات حول القيمة المتوسطة. إذا كانت القيمة المتوسطة قريبة جداً من جميع القيم فإن التباين يكون منخفضاً، بينما إذا كانت القيم متشتتة بشكل كبير يكون التباين عالياً. ومن خلال حساب التباين بهدف الاستغناء عن السمات ذات التباينات الصفرية أي السمات التي لم تبدي أي تغير على طول مجموعة البيانات ومن أجل كل السجلات في مجموعة البيانات.

العلاقة الرياضية التي تعبر عن التباين:

$$Cov(x)=\frac{1}{n}\sum_{i=1}^{n}(xi-x) \qquad (3)$$

حيث: x : المتوسط الحسابي لقيم xi ، n: عدد العينات.

يتم بالاعتماد على العلاقة السابقة (3) من خلال التعليمة البرمجية var(x) في لغة البايثون حيث x هي العمود المعبر عن السمة المختارة في كل مرة، فكانت النتيجة كما هو موضح بالجدول (2):

جدول (2) : حساب التباين من أجل كل سمة

MFCC(1)=10789.108	MFCC(2)=1693.25	MFCC(3)=327.74	MFCC(4)=579.58
MFCC(5)=162.30	MFCC(6)=157.61	MFCC(7)=58.67	MFCC(8)=53.81
MFCC(9)=53.49	MFCC(10)=30.77	MFCC(11)=94.91	MFCC(12)=50.03
MFCC(13)=59.70	MFCC(14)=45.23	MFCC(15)=23.95	MFCC(16)=23.63
MFCC(17)=14.54	MFCC(18)=21.95	MFCC(19)=16.33	MFCC(20)=21.16
MFCC(21)=24.42	MFCC(22)=21.94	MFCC(23)=23.25	MFCC(24)=29.28
MFCC(25)=27.89	MFCC(26)=25.34	MFCC(27)=17.98	MFCC(28)=17.86
MFCC(29)=14.92	MFCC(30)=17.95	**MFCC(31)=0**	MFCC(32)=17.41
MFCC(33)=16.07	MFCC(34)=21.89	MFCC(35)=31.22	MFCC(36)=32.18
MFCC(37)=0	MFCC(38)=31.90	MFCC(39)=25.86	MFCC(40)=15.77
MFCC(41)=12.64	MFCC(42)=9.85	MFCC(43)=9.63	**MFCC(44)=0**
MFCC(45)=7.46	MFCC(46)=7.21	MFCC(47)=5.80	MFCC(48)=7.05
MFCC(49)=7.01	MFCC(50)=6.08	MFCC(51)=5.36	MFCC(52)=4.62
MFCC(53)=3.89	MFCC(54)=3.89	MFCC(55)=3.34	MFCC(56)=3.47
MFCC(57)=3.38	MFCC(58)=3.28	MFCC(59)=3.20	

مما سبق نجد أن السمات 31، 37، 44 لها تباين صفري أي لا تتغير قيمتها من أجل كل السجلات في مجموعة البيانات وبالتالي يمكن حذفهم مما يقلل من حجم مجموعة البيانات دون التأثير على دقة التعرف وأصبح عدد السمات بعد هذه المرحلة 56 سمة.

7-3 بيانات التدريب وبيانات التحقق:

يتم تدريب النموذج على مجموعة بيانات من التسجيلات الصوتية حيث تعمل بيانات التدريب هذه كأساس للنموذج لمعرفة الأنماط والعلاقات بين أشكال الموجات الصوتية والكلمات المنطوقة التي تمثلها، ولتقييم قدرة تعميم النموذج ومنع التجهيز الزائد يتم استخدام مجموعة بيانات منفصلة للتحقق من الصحة. ثم يتم تقييم الأداء في مجموعة البيانات هذه في نهاية كل فترة باستخدام وظيفة رد الاتصال المخصصة (CallbackEval). يحسب رد الاتصال هذا معدل خطأ الكلمات (WER)، وهو مقياس يحدد الفرق بين تنبؤات النموذج والنسخ الحقيقية وبالإضافة إلى ذلك، يتم عرض تنبؤات العينة للتقييم النوعي.

7-4 تقييم أداء النموذج:

تم تدريب النموذج ل 50 حقبة زمنية. استغرقت كل حقبة ما يقرب من 27 إلى 29 دقيقة باستخدام وحدة معالجة الرسومات GeForce RTX 2080 Ti. وظهرت النتائج كما في الجدول (3):

جدول (3) :معدل الخطأ الكلمات والخسارة خلال كل حقبة

Epoch	WER	Val-loss
1	**1.0000**	**272.9774**
2	**0.9829**	**267.4091**
3	0.9657	261.8408
4	0.9486	256.2725
5	0.9314	250.7042
6	0.9143	245.1358
7	0.8971	239.5675
8	0.8800	233.9992
9	0.8629	228.4309
10	0.8457	222.8626
11	0.8286	217.2943
12	0.8114	211.7260
13	0.7943	206.1577
14	0.7771	200.5893
15	0.7600	195.0210
16	0.7429	189.4527
17	0.7257	183.8844
18	0.7086	178.3161
19	0.6914	172.7478
20	0.6743	167.1795
21	0.6571	161.6112
22	0.6400	156.0428
23	0.6229	150.4745
24	0.6057	144.9062
25	0.5886	139.3379

Epoch	WER	Val-loss
26	0.5714	133.7696
27	0.5543	128.2013
28	0.5371	122.6330
29	0.5200	117.0647
30	0.5029	111.4963
31	0.4857	105.9280
32	0.4686	100.3597
33	0.4514	94.7914
34	0.4343	89.2231
35	0.4171	83.6548
36	0.4000	78.0865
37	0.3829	72.5182
38	0.3657	66.9498
39	0.3486	61.3815
40	0.3314	55.8132
41	0.3143	50.2449
42	0.2971	44.6766
43	0.2800	39.1083
44	0.2629	33.5400
45	0.2357	27.9717
46	0.2186	22.4033
47	0.2114	16.8350
48	0.1693	11.2667
49	**0.1671**	**5.6984**
50	**0.1600**	**0.1301**

في الحقبة الأولى، بدأ النموذج للتو في التعلم من البيانات. لا يستطيع التعرف على الكلام بدقة عالية، ومعدل خطأ الكلمة مرتفع جداً (1.0000). ومع ذلك فإن النموذج قادر على التعلم بسرعة، وينخفض معدل الخطأ في الكلمات بشكل ملحوظ في الحقبة الثانية.

إن معدل الخطأ الكلمات في الحقبة الثانية هو 0.9829. ويعد هذا تحسناً كبيراً مقارنة بالحقبة الأولى، ويظهر أن النموذج مستمر في التعلم من البيانات. ولا يزال النموذج يرتكب الأخطاء لكنه يرتكب أخطاء أقل مما كان عليه من قبل. كما نلاحظ أن معدل الخطأ في الكلمات في الحقبة الأخيرة هو 0.1600. ويعد هذا معدل خطأ منخفضاً جداً ويشير إلى أن النموذج قادر على التعرف على الكلام بدقة شديدة. وقد تعلّم النموذج التعرف والتمييز بين الكلمات المختلفة، وأصبح قادراً على التنبؤ بالكلمات الصحيحة بدرجة عالية من الدقة.

يتحسن أداء النموذج بشكل مطرد على مدار التدريب. حيث يتناقص معدل خطأ الكلمات مع كل فترة ويتمكن النموذج من نسخ الصوت بدقة أكبر. وكان أداء النموذج جيد جداً بشكل خاص في الحقبة الأخيرة، حيث حقق معدل خطأ منخفض جداً في الكلمات يبلغ 16%.

تتناقص قيمة الخسارة بشكل مطرد على مدار التدريب، وتصل إلى قيمة منخفضة تبلغ 0.1301 في الفترة الأخيرة. ويشير هذا إلى أن النموذج قادر على تعميم البيانات غير المرئية بشكل جيد، كما أنه قادر على التعرف على الكلمات بدقة على التسجيلات الجديدة.

8- الاستنتاجات والتوصيات

تم في هذا البحث تصميم نموذج قادر للتعرف التلقائي على الكلام بالاعتماد على تقنيات التعلم العميق من خلال تدريبة على قاعدة بيانات كبيره واختباره على مجموعة بيانات الاختبار، ومن خلال دراستنا السابقة نجد:

- أعطى نموذج التعلم العميق المقترح أفضل نتيجة من حيث دقة التعرف حيث بلغ معدل الخطأ في الكلمات 16% بالمقارنة مع النماذج الأخرى يعود ذلك الى البنية الهجينة التي يتمتع بها النموذج المقترح وحصولنا على معدل خطأ الكلمات قليل جداً.
- حسنت خطوات المعالجة الأولية وخطوات زيادة البيانات دقة التعرف على الكلمات وقدرتنا على تعميم النموذج في البيئات المختلفة.
- يجدر بالملاحظة أن استخدام معدل خطأ الكلمات القليل يشير إلى أن النموذج يتمتع بقدرة عالية على تفسير التسجيلات الصوتية والتعرف على الكلمات بدقة. ويمكن أن يكون لهذا تأثير إيجابي على التطبيقات الفعلية حيث يتم تحويل الكلام إلى نص أو فهم أوامر الصوت بشكل أفضل.

التوصيات:

- مواصلة تحسين النموذج: يُنصح بمواصلة العمل على تحسين النموذج القائم على تقنيات التعلم العميق، سواء من خلال تحسين البنية الهجينة للنموذج أو استخدام تقنيات التعلم العميق الأحدث.
- استخدام تقنيات المعالجة الأولية وزيادة البيانات: يجب الاستمرار في تحسين وتطوير عمليات المعالجة الأولية للبيانات وتقنيات زيادة البيانات.
- استكشاف التطبيقات الفعلية: يُنصح بدراسة واستكشاف التطبيقات الفعلية للنموذج المطور، مثل تحويل الكلام إلى نص أو فهم أوامر الصوت. قد يكون لاستخدام معدل خطأ الكلمات القليل تأثير إيجابي على تلك التطبيقات، حيث يمكن أن يترجم ذلك إلى دقة أفضل وتجربة مستخدم أفضل.

المراجع

[1] Chaturvedi, S. K., Saxena, D., Taku, R., Saluja, V. K., Bhattarai, S., & Pradesh, A. (2023) PHONETICS AND COMMUNICATION: BRIDGING THE GAP BETWEEN SPEECH AND UNDERSTANDING.

[2] 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.

[3] Dhakal, P., Damacharla, P., Javaid, A. Y., & Devabhaktuni, V. (2019). A near real-time automatic speaker recognition architecture for voice-based user interface. Machine learning and knowledge extraction, 1(1), 504-520.

[4] Deshmukh, A. M. (2020). Comparison of hidden markov model and recurrent neural network in automatic speech recognition. European Journal of Engineering and Technology Research, 5(8), 958-965.

[5] Ramadan, R., & Yadav, K. (2021). Nonlinear acoustic noise cancellation based automatic speech recognition system (NANC-ASR) with convolutional neural networks. International Journal of Speech Technology.

[6] Veisi, H., & Haji Mani, A. (2020). Persian speech recognition using deep learning. International Journal of Speech Technology, 23, 893-905.

[7] https://keithito.com/LJ-Speech-Dataset/

[8] https://www.kaggle.com/datasets/ejlok1/cremad

[9] https://www.openslr.org/12

[10] http://www.speech.sri.com/

[11] Lu, L., Meng, Z., Kanda, N., Li, J., & Gong, Y. (2020). On minimum word error rate training of the hybrid autoregressive transducer. arXiv preprint arXiv:2010.12673.

[12] Lee, J., & Watanabe, S. (2021, June). Intermediate loss regularization for ctc-based speech recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6224-6228). IEEE.

[13] Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 10, 122136-122158.