مجلة جامعة دمشق للعلوم الهندسية
المجلد40 العدد(4) – 2024: 389-411

Damascus University Journal for engineering sciences
Vo 40 number (4) -2024: 389-411

بناء نظام تعرف الكلمات المفتاحية باللغة العربية باستخدام التعلم الخاضع للإشراف الذاتي

أسامة ابراهيم ديب[1]* أميمة الدكاك[2] آصف جعفر[3]

*[1]. طالب دراسات عليا (دكتوراه) مهندس، المعهد العالي للعلوم التطبيقية والتكنولوجيا، هندسة الاتصالات.
E-mail: osama.deeb@hiast.edu.sy

[2]. دكتورة، أستاذة، المعهد العالي للعلوم التطبيقية والتكنولوجيا، دكتوراه في معالجة الإشارة الكلامية والذكاء الصنعي، E-mail: oumayma.dakkak@hiast.edu.sy

[3]. دكتور، أستاذ، المعهد العالي للعلوم التطبيقية والتكنولوجيا، دكتوراه في الذكاء الصنعي ومعالجة الإشارة
E-mail: assef.jafar@hiast.edu.sy

الملخص:

تعرض هذه الورقة البحثية دراسةً لاستخدام نماذج التمثيل السياقي المدرَّبة باستخدام التعلُّم الخاضع للإشراف الذاتي في مسألة تعرُّف الكلمات المفتاحية باللغة العربية، بهدف تقليل كمية المعطيات اللازمة للتدريب، مع الحفاظ على دقة عالية في التعرُّف. تم استخدام نموذج Hidden Unit Bidirectional Encoder Representations from Transformers (HuBERT) لاستخراج التمثيل السياقي للإشارة الكلامية المدرب مسبقاً على معطيات باللغة العربية وبناء نموذج رأسي خاص بمسألة التعرُّف، تبعه إجراء معايرة دقيقة للنموذج الكلي باستخدام المدوَّنة Arabic Speech Command، وتم إجراء سلسلة من التجارب بهدف تحديد الحد الأدنى من عينات التدريب اللازمة لتحقيق دقة تعرُّف معينة.

حققت البنية المقترحة دقة تعرُّف تجاوزت 98.5% باستخدام عشر عينات تدريب لكل كلمة فقط، وتجاوزت 99.7% بزيادة العدد إلى 11 عينة تدريب أو أكثر. تم اختبار البنية على اللغة الانكليزية أيضاً، بهدف المقارنة، وحققت نتائج مماثلة من حيث الدقة وعدد العينات اللازمة للتدريب. تبين النتائج فعالية التعلم الخاضع للإشراف الذاتي في مسألة تعرُّف الكلمات المفتاحية باللغة العربية فيما يتعلق بتقليل عدد العينات اللازمة للتدريب، وإمكانية استخدامه في تطبيقات أوسع لمعالجة الإشارة الكلامية.

الكلمات المفتاحية: التمثيل السياقي، التعلم الخاضع للإشراف الذاتي، تعرُّف الكلمات المفتاحية، مدوَّنة.

تاريخ الايداع: 2023/10/26
تاريخ القبول: 2024/4/21

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

Building a Keyword Spotting Model for Arabic Language Using Self-Supervised Learning Approach

Osama Ibrahim Deeb[*1] Oumayma Al-Dakkak[2] Assef Jafar[3]

[*1.] PhD student, Eng, Higher Institute for Applied Sciences and Technology (HIAST), telecommunication engineering. E-mail: osama.deeb@hiast.edu.sy

[2]. Professor, Dr, Higher Institute for Applied Sciences and Technology (HIAST), Speech Signal Processing and AI. E-mail: oumayma.dakkak@hiast.edu.sy

[3]. Professor Dr, Higher Institute for Applied Sciences and Technology (HIAST), AI and Signal Processing. E-mail: assef.jafar@hiast.edu.sy

Abstract:

This research paper presents a comprehensive investigation into the efficiency of using contextual representation models trained via self-supervised learning for keyword spotting (KWS) in the Arabic language, in view of reducing the amount of data required for training, while maintaining high accuracy in KWS. We employed Hidden Unit Bidirectional Encoder Representations from Transformers (HuBERT), a pre-trained model on Arabic data for extracting the contextual representation of the speech signal and developed a head model for the KWS downstream task. This head model was fine-tuned using the Arabic Speech Command dataset, and multiple experiments were conducted to ascertain the minimum number of training samples required to attain a specific level of accuracy.

Remarkably, using only ten training samples per word, the achieved detection accuracy exceeded 98.5%, and by increasing the number to more than 11 training samples, the accuracy increased to 99.7%. The performance of the model was evaluated on English language data and obtained similar outcomes regarding accuracy and the number of training samples needed for training. The results demonstrate the effectiveness of self-supervised learning for the KWS task in Arabic regarding the reduction of required training samples and suggest the potential for broader applications in speech processing.

Keywords: Contextual Representation, Self-Supervised Learning, Keyword Spotting, HuBERT, Dataset.

Received: 26/10/2023
Accepted: 21/4/2024

المقدمة (Introduction):

تعرُّف الكلمات المفتاحية Keyword Spotting (KWS) هي عملية مراقبة مستمرة للكلام بهدف التحسس إلى ورود كلمة أو مجموعة من الكلمات المحددة مسبقاً ضمن سياق الكلام (Sainath and Parada 2015)، وتكتسب هذه المسألة أهميةً متزايدةً مع التطور السريع في مجال الأجهزة الذكية وأنترنت الأشياء، حيث يصبح التحكم الصوتي بالأجهزة المحيطة غايةً تسعى إليها معظم الشركات الكبرى لتحقيق المنافسة نظراً لما تؤمنه من سهولة في الاستخدام وخصوصاً للأشخاص الذين يعانون من بعض الإعاقات (Chen, Parada and Heigold 2014).

إن استخدام تقنيات التعرف الآلي على الكلام Automatic Speech Recognition (ASR) لتحويل الأوامر الصوتية إلى نص مكتوب ومحاولة تفسيرها والتفاعل معها تعاني من العديد من الصعوبات، أهمها: أن هذه الأنظمة بحاجة إلى موارد حسابية ضخمة لتحقيق أداءٍ مقبول، وبالتالي يصعب تشغيلها على الأجهزة الذكية في الوقت الراهن، إضافة إلى ذلك، فإن فكرة المعالجة السحابية الدائمة مستبعدة لأسباب تتعلق بالأمان وانتهاك الخصوصية والكم الهائل من تبادل المعطيات بين الجهاز والسحابة (Tang and Lin 2018). من هنا تظهر أهمية نظم تعرُّف الكلمات المفتاحية كوسيلة فعالة ذات متطلبات حسابية قليلة، يمكن الاستعانة بها لفهم وتنفيذ أوامر محددة، أو كأداة لإطلاق خدمات سحابية أخرى بناءً على كلمات محددة، وهذه المنهجية تستخدمها عملياً بعض الشركات، مثل إطلاق خدمة البحث الصوتي في غوغل عند نطق كلمة "Okay google" في الأجهزة الذكية العاملة بنظام أندرويد، أو التحدث إلى المساعد الافتراضي في أمازون "Alexa"، أو المساعد الافتراضي في أنظمة آبل "Siri" (Sainath and Parada 2015).

لا يقتصر استخدام KWS على تطبيقات الأجهزة الذكية، إذ يمكن استخدامها في عمليات البحث السريع ضمن الملفات الصوتية المخزنة، وعمليات التصنيف، والوسم الآلي للملفات الصوتية، فضلاً عن استخدامها في مجال مراقبة الاتصالات الهاتفية وتعقب المكالمات ذات المحتوى الإجرامي (Salhab and Harmanani 2023).

تنتمي التقنيات المستخدمة في تعرُّف الكلمات المفتاحية إلى منهجين أساسيتين: المنهج التوليدي (generative) والمنهج التمييزي (discriminative). في المقاربة التوليدية –المسماة أحياناً المقاربة المبنية على النموذج (model-based)– يتعلم النظام نموذج الكثافة الاحتمالية المشتركة $\mathbf{p}(\mathbf{x},\mathbf{y})$، للإشارة الصوتية $\mathbf{x}$ و الكلمة المفتاحية $\mathbf{y}$، ثم يقوم باستخدام هذا النموذج لحساب الكثافة الاحتمالية الشرطية $\mathbf{p}(\mathbf{y}/\mathbf{x})$ بالاعتماد على قاعدة Bayes، ومن ثم يتم اعتماد الكلمة المفتاحية y ذات الاحتمال الأكبر تبعاً لهذا النموذج. في حين يتركز الهدف الأساسي للمقاربات التمييزية على تعظيم احتمالية كشف سلسلة من المقاطع الصوتية تمثل الكلمة المرغوب تعرُّفها وذلك بالاعتماد على مجموعة من السمات المستخرجة من الإشارة الصوتية الواردة (Tabibian, Shokri, et al. 2011).

تعتمد المقاربات التوليدية بشكل أساسي على نماذج ماركوف المخفية Hidden Markov Models (HMM)، وتقسم إلى ثلاث مجموعات رئيسية هي: النماذج القائمة على نمذجة الكلمة ككل whole-word modeling وأحياناً تسمى Acoustic KWS، والنماذج القائمة على البحث على مستوى الصوتيم Phonetic Search KWS، والنماذج القائمة على نظم تعرف الكلام للسلاسل الطويلة LVCSR-Based KWS (Large Vocabulary Continuous Speech Recognition) (Tabibian, Shokri, et al. 2011). لكلٍ من هذه الطرق مجالٌ أمثلي للاستخدام، ولكنها تعاني مجتمعةً من مشكلة بنيوية تتعلق بالهدف الأساسي لعملية التدريب، حيث يهدف

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

تـدريب نمـاذج HMM إلـى تعظـيم الأرجحيـة (likelihood) للمقاطع الصوتية التي يتم التدريب عليها، ولكنه لا يقدم أي ضمانات فيما يتعلق بالأداء في مرحلة الاختبار؛ بمعنى أنه يسعى خلال مرحلة التدريب إلى بناء أفضل النماذج للكلمات المطلوبة، ولكنه لا يضمن تمييز هذه الكلمات عن الكلمات غير المفتاحية في مرحلة الاختبار، وفي كثير من الأحيان، يمكن أن يحدث خلط مع النموذج الممثل للكلمات غير المفتاحية ويتسبب ذلك بتجاهل الكلمة المفتاحية الواردة على دخل النظام (Keshet and Bengio 2009).

أما فيما يتعلق بالنظم القائمة على المقاربات التمييزية فيمكن تقسيم عملها إلى مرحلتين: مرحلة استخراج السمات Feature Extraction، ومرحلة التصنيف Classification. في المرحلة الأولى يتم استخراج بعض السمات التمييزية من المقاطع الكلامية المدخلة لتُستخدم في نمذجة معامل الثقة ومكان الورود للكلمات المفتاحية في المقاطع المدخلة، وفي المرحلة الثانية يتم استخدام خوارزمية تصنيف للفصل بين الكلمات المفتاحية والكلمات غير المفتاحية. تأتي أفضلية المقاربات التمييزية على المقاربات التوليدية من حقيقة أن النموذج في طور التدريب يتعلم الفصل بين الألفاظ المطلوبة والألفاظ المغايرة، وبالتالي كلما زاد تدريب النموذج كلما أصبح أكثر قدرةً على الفصل والتمييز (Tabibian, Akbari and Nasersharif 2018).

تقسم المقاربات التمييزية المطبقة في تعرُّف الكلمات المفتاحية إلى مجموعتين أساسيتين: تعرُّف الكلمات المفتاحية باستخدام آلة الشعاع الداعم Support Vector Machine (SVM)، وتعرُّف الكلمات المفتاحية باستخدام الشبكات العصبونية Neural Networks (NN). إن SVM فعالة في حالات الفصل الثنائي ولكن أداءها يتراجع بشكل كبير في حال الفصل المتعدد مما يجعلها فعالة في الحالات التي يُطلب فيها الكشف عن ورود الكلمة فقط دون الاهتمام بالتعرف عليها. أما الشبكات العصبونية بمختلف أنواعها فقد أثبتت الدراسات المرجعية أنها الأداة الأفضل لمعالجة هذه المسألة، لأنها قادرة على نمذجة الإشارة الكلامية بشكل يراعي الكثير من خصوصياتها وبالتالي تكون النماذج المبنية باستخدامها قادرة على تجاوز المشاكل المرتبطة بطبيعة الإشارة الكلامية (اختلاف سرعة الكلام، النبرة، الحالة النفسية، ..... إلخ) (Tabibian, Shokri, et al. 2011).

تم في هذا البحث استخدام أحدث المفاهيم والتقنيات في مجال الشبكات العصبونية والتعلم العميق، وأُعطي اهتمام خاص للتعلم الخاضع للإشراف الذاتي واستخدام تقنيات نقل التعلُّم للتغلب على المشاكل المرتبطة باستخدام اللغة العربية في هذا المجال، وبشكلٍ خاص مشكلة نقص معطيات التدريب.

تتضمن المقالة، بالإضافة إلى هذه المقدمة، عرضاً لأهم الدراسات المرجعية الحديثة المتعلقة بمسألة البحث في المقطع الثاني. ثم يُفصّل المقطع الثالث المنهجية المتَّبعة في هذا البحث لحل المسألة مع توضيح جميع المكونات التي تم استخدامها. ويعرض المقطع الرابع النتائج التي تم التوصل إليها قبل أن تُعرض الاستنتاجات والآفاق المستقبلية في المقطع الخامس والأخير.

2. الدراسات المرجعية (LiteratureReview):

إن الدراسات المرجعية المتعلقة بمسألة تعرُّف الكلمات المفتاحية باللغة العربية شحيحة نوعاً ما، على الرغم من سعة انتشار اللغة العربية ومكانتها المتقدمة من حيث عدد المتكلمين. قدم (Awaid, Fawzi and Kandil 2014) نموذجاً لتعرُّف الكلمات المفتاحية لاستخدامه في البحث الصوتي في المقاطع الصوتية التي يتوفر لها تدوين نصي كالقرآن الكريم، حيث اعتمدت الخوارزمية المقدمة على المزج بين البحث النصي والبحث الصوتي عن المقطع المطلوب ونجحت بتحقيق دقة كشف بلغت 97%، إلا أن هذه الدقة تنخفض إلى 84% عند اختبارها على مقاطع صوتية لا يتوفر تدوين نصي لها. كما قدم (Salhab and Harmanani 2023) نموذجهم المسمى AraSpot لتعرُّف الكلمات المفتاحية باللغة العربية، واستخدموا فيه البنية ConformerGRU التي تستخدم تقنية الانتباه

المتعدد الرؤوس Multi-Head Attention من أجل لحظ الترابط على المدى القصير والطويل في الكلام، واستخدموا في تدريبه المدونة Arabic Speech Command، واحتاجوا إلى إجراء عمليات زيادة المعطيات data augmentation وتوليد عينات للتدريب باستخدام أدوات توليد الكلام من النصوص text to speech (TTS) من أجل تأمين كمية معطيات كافية لتدريب الشبكة وتحقيق التقارب، واستطاعوا تحقيق دقة كشف بلغت 99.59%.

كذلك توجد مجموعة من الدراسات المرجعية تحت مسمى تعرُّف الكلمات العربية "Arabic Word Spotting" ولكنها تهتم بتعرُّف الكلمات باللغة العربية في النصوص وليس في الكلام. فمثلاً قدم (Cheikh Rouhou, Kessentini and Kanoun 2019) نموذجاً هجيناً استخدموا فيه نماذج ماركوف المخفية HMM والشبكات العصبونية ذات التوصيل الكامل والتي تسمى أحياناً بالشبكات العصبونية العميقة Deep Neural Networks (DNN) لبناء نظام لتعرُّف الكلمات العربية في النصوص المكتوبة بخط اليد. استخدم (Brik, et al. 2014) تحويل Curvelet لتوليد صور أكثر دقة للكلمات من أجل استخدامها في نظام تعرُّف الكلمات العربية في الوثائق التاريخية القديمة التي تعاني من بعض التلف، حيث طُبق التحويل المذكور على عينات التدريب في المدونة وساهم في تحسين الأداء عند تجربه على وثائق ورقية قديمة. كما قدم (Fathallah, El-Yacoubi and Ben Amara 2023) نموذجاً لتعرُّف الكلمات العربية في الوثائق المصورة واستخدموا فيه تقنية نقل التعلم للاستفادة من معارف سابقة مطبقة على وثائق مكتوبة بخط اليد للغتين العبرية والإنكليزية ومن ثم استخدامها للغة العربية، واختبر النموذج على قاعدة المعطيات Arabic VML-HD وأظهرت النتائج تفوق هذا النموذج على النماذج المماثلة في وقته. يوجد عدة دراسات مماثلة ولكن القاسم المشترك بينها أنها تهتم بتعرُّف الكلمات العربية في النصوص المخزنة على شكل صور، وهذا النوع من الأعمال يندرج في إطار معالجة الصورة، وهو بعيد بشكل كبير عن موضوع معالجة الإشارة الكلامية الذي يهتم به هذا البحث.

يوجد العديد من الدراسات المرجعية حول موضوع تعرُّف الكلمات المفتاحية ولكن بلغات أخرى غير العربية، حيث كانت المحاولات الأولى للتصدي للمسألة كما ذُكر في المقدمة مبنية على أساس المقاربات التوليدية باستخدام سلاسل ماركوف، وقد واجهت هذه الطرق العديد من المشاكل وعجزت عن إيجاد الحلول لها. مع بداية الألفية الثالثة، وظهور المعالجات ذات القدرات الحسابية الكبيرة أصبح استخدام الشبكات العصبونية أمراً قابلاً للتطبيق عملياً؛ الأمر الذي فتح آفاقاً جديدةً أمام الباحثين للاستفادة من الميزات التي توفرها هذه الشبكات من أجل تلافي المشاكل التي عجزت نماذج ماركوف عن حلها.

بين (Chen, Parada and Heigold 2014) تفوق أداء نظم التعرُّف المبنية باستخدام الشبكات العصبونية ذات التوصيل الكامل Deep Neural Networks (DNN) على النظم المبنية باستخدام نماذج ماركوف المخفية HMM. حاول (Sainath and Parada 2015) بناء نموذج قادرٍ على العمل ضمن موارد محدودة من حيث استهلاك الذاكرة والقدرات الحسابية، ولهذا السبب استخدم الشبكات العصبونية الالتفافية Convolutional Neural Networks (CNN) لقلة عدد المتحولات الذي تحتاجه بالنسبة للشبكات ذات التوصيل الكامل. بيّنت هذه الدراسة أن استخدام الشبكات الالتفافية CNN يتفوق على أداء الشبكات (DNN) من ناحية دقة الكشف وتقليل عدد المتحولات. حاول (Arık, et al. 2017) المزج بين الشبكات الالتفافية CNN، والشبكات العودية (التكرارية) Recurrent Neural Networks (RNN) بحيث يتمتع النموذج المقترح بمزايا الشبكتين؛ إذ تمتاز شبكات CNN بالقدرة على لحظ الترابط القصير الأمد في الإشارة الكلامية، في حين تتفوق شبكات RNN في لحظ الترابط الطويل الأمد. وقد حقق هذا المزج تحسناً كبيراً في الأداء ظهر من خلال الممانعة الكبيرة للضجيج التي أبداها النظام.

استفاد الباحثون في هذا المجال من النجاحات التي حققتها بعض الشبكات العصبونية في مجالات أخرى وحاولوا تطبيقها في مجال KWS، فحاول (Tang and Lin 2018) إدخال التعلم العميق المتبقي (deep residual learning) –الذي استخدم في شبكة Residual Network(ResNet) وحقق قفزةً نوعية في مجال الرؤية الحاسوبية Computer Vision (CV)– على مسألة KWS وذلك من أجل دراسة أثر عمق الشبكة على أداء النظام. خلصت الدراسة إلى أن تطبيق تقنية التعلم المتبقي أسهم إلى حد كبير في تحسين الأداء من خلال القدرة على زيادة عمق الشبكة دون ظهور آثار سلبية على عملية تعليمها. استفاد (Coucke, et al. 2019) من النجاح الذي حققته فكرة الترابط الممدد Dilated Convolution في مجال نمذجة السلاسل وإسقاطه على مسألة تعرُّف الكلمات المفتاحية. كذلك استفاد من فكرة توابع التفعيل ببوابات Gated Activations والروابط المتبقية Residual Connections التي طرحت في نموذج WaveNet[1] المصمم من قبل شركة Google والمستخدمة في توليد الكلام من النص (Text To Speech TTS) في تصميم نظام للتعرف على الكلمة المفتاحية "Hey Snips"، وقام الباحث بمقارنة الأداء مع الشبكات العودية التقليدية والشبكات الالتفافية التقليدية فكانت دقة النموذج أفضل من كلتا الحالتين وخصوصاً في حال وجود الضجيج.

افترضت معظم الدراسات السابقة وجود كمية كافية من المعطيات من أجل تدريب الشبكات العصبونية. في الواقع، تعتبر مسألة توفر معطيات التدريب عنق الزجاجة لكثير من المسائل التي تستخدم الشبكات العصبونية في حلها، ومسألة تعرُّف الكلمات المفتاحية واحدةٌ من تلك المسائل؛ إذ تحتاج بعض الشبكات إلى 4000 عينة لكل كلمة من أجل تدريبها (Lin, et al. 2020)، وعند تغيير قائمة الكلمات المطلوب تعرُّفها يجب إعادة تدريب النظام من جديد ليتعلم الكشف عن الكلمات الجديدة، بمعنى آخر، لا بدَّ من توفير كمية من التسجيلات الصوتية لقائمة الكلمات الجديدة من أجل إعادة التدريب، وعندما تتكرر هذه الحالة بشكلٍ دائمٍ يصبح هذا العمل مضنياً وغير مجدٍ من الناحية العملية. دفع هذا الأمر الباحثين إلى الاستعانة بتقنيات كنقل التعلم Transfer Learning والتعلم الخاضع للإشراف الذاتي Self-Supervised Learning من أجل تقليل عدد العينات اللازمة للتدريب، حيث قدم (Seo, Oh and Jung 2021) نموذجاً للكشف عن الكلمات المفتاحية باللغة الكورية يعتمد على مبدأ نقل التعلم من خلال استخدام نموذج wav2vec 2.0[2] مدرّب مسبقاً على قاعدة معطيات باللغة الانكليزية، ونموذج wav2vec2.0 هو نموذج يستخدم مبدأ التعلم الخاضع للإشراف الذاتي من أجل استخلاص تمثيل للإشارة الكلامية يأخذ بالاعتبار ورود كل كلمة ضمن سياق الجملة ويسمى التمثيل السياقي contextual representation. نجح النموذج المقدّم بتحقيق دقة كشف بلغت 95% من خلال إضافة نموذج رأسي خاص بمسألة التعرُّف وإعادة تدريب النموذج الكلي بواسطة قاعدة معطيات تتضمن 20 لفظاً لكل كلمة مفتاحية باللغة الكورية، واستطاع بذلك تقليل عدد عينات التدريب مع الحفاظ على دقة جيدة.

قدم (Yang, et al. 2021) بحثاً عرض فيه استخدام مجموعة من النماذج المدربة وفق تقنية التعلم الخاضع للإشراف الذاتي مثل wav2vec 2.0 و HuBERT[3] وغيرها في مسائل تتعلق بمعالجة الإشارة الكلامية مثل تعرف الصوتيمات Phoneme Recognition (PR)، والتعرف الآلي على الكلام Automatic Speech Recognition (ASR)، وتحديد هوية المتكلم

[1] WaveNet: شبكة عصبونية عميقة استخدمت في مسألة توليد الكلام (2016)

[2] Wav2vec 2.0: نموذج للتمثيل السياقي للإشارة الكلامية يستخدم تقنية التعلم الخاضع للإشراف الذاتي عبر تكمية السمات المضمَّنة للمقاطع الصوتية.

[3] HuBERT: Hidden Unit Bidirectional Encoder Representations from Transformers.

Speaker Identification (SID)، والتحقق من هوية المتكلم Automatic Speaker Verification (ASV)، وKWS وغيرها. قامت الفكرة الأساسية للبحث على استخدام نسخ مدربة مسبقاً باستخدام معطيات باللغة الانكليزية من هذه النماذج وإضافة نموذج رأسي يناسب المسألة المطلوبة ومن ثم إعادة تدريب النموذج الرأسي فقط، حيث يمكن اعتبار النتائج المعروضة في هذه الدراسة كمرجع للمقارنة بين أداء هذه النظم.

تم في هذا البحث استخدام نموذج HuBERT مدرب مسبقاً ضمن البنية المقترحة لتعرُّف الكلمات المفتاحية باللغة العربية.

ونموذج HuBERT وهو نموذج يعتمد مبدأ التعلم الخاضع للإشراف الذاتي، ويستخدم لاستخراج التمثيل السياقي للإشارة الكلامية، ويعتمد في بنيته بشكلٍ أساسي على المحولات (Transformers) التي تستخدم تقنية الانتباه متعدد الرؤوس (Multi-Head Attention). استُخدم هذا النموذج في العديد من الأبحاث الخاصة باللغة العربية. حيث استخدم (Mohamed and A. Aly 2021) هذا النموذج في نظام للتعرف على المشاعر في المحادثات باللغة العربية، كذلك استخدم في الدراسة (Almutairi and Elgibreen 2023) لبناء نظام لكشف تزييف الكلام العربي المولد باستخدام الآلة عن طريق تقنيات التزييف العميق Deepfake. قدم (Waheed, et al. 2023) نظاماً لتعرف الكلام باللغة العربية يعمل على عدة لهجات عربية كالمصرية والمغربية بالإضافة إلى اللغة العربية الفصحى الحديثة(MSA) Modern Standard Arabic واستخدم في بناءه نموذج HuBERT مدرباً مسبقاً. على حد معرفتنا ومن خلال الدراسات المرجعية التي اطلعنا عليها فإنه لم يتم استخدام هذا النموذج في مسألة تعرُّف الكلمات المفتاحية للغة العربية من قبل.

جرى التركيز في هذا البحث على مسألتين أساسيتين هما: بناء نظام تعرُّف باللغة العربية، وأن يكون هذا النظام قابلاً للتدريب باستخدام عدد محدود من العينات مما يسهم في تقليل الكلفة اللازمة لتغيير قائمة الكلمات المطلوب تعرُّفها. بعد الاطلاع على العديد من الدراسات والأبحاث في هذا المجال، تبين أن الحل الأنسب يكون باستخدام نماذج مدربة مسبقاً على استخراج التمثيل السياقي للإشارة الكلامية ثم بناء نموذج رأسي بتعقيد مناسب يستخدم هذا التمثيل السياقي لتحقيق وظيفة تعرُّف الكلمات المفتاحية، وتدريب هذا النموذج باستخدام عدد قليل من عينات التدريب باللغة العربية.

3- مواد البحث وطرائقه (Materials and Methods):

يتضمن هذا القسم فقرةً تتناول مواد البحث وأخرى لطرائقه.

3-1- مواد البحث:

تم في هذا الجزء شرح أهم المفاهيم التي استُخدمت في هذا البحث وهي: التعلم الخاضع للإشراف الذاتي، وآلية نقل التعلم، ونموذج HuBERT، كذلك تم تقديم توصيف للمدونتين المستخدمتين في التدريب والاختبار وهما: Google Speech Command وArabic Speech Command.

3-1-1- التعلم الخاضع للإشراف الذاتي Self-Supervised Learning (SSL):

يتم تدريب الشبكات العصبونية في التعلم الخاضع للإشراف (supervised learning) باستخدام قواعد معطيات موسومة (labeled dataset)، حيث تحتوي القاعدة بالإضافة إلى عينات التدريب على وصفٍ لكل عينة يناسب المهمة التي تستخدم القاعدة من أجلها، ويتم استخدام الوسوم من أجل حساب تابع الخسارة الذي يجري تعديل أوزان الشبكة على أساسه أثناء عملية التدريب. أما في التعلم الخاضع للإشراف الذاتي، فيتم استخدام معطيات غير موسومة في التدريب، حيث يتم بدايةً تعريف مهمة أولية أو ذريعة (pretext) يتم اشتقاقها من المعطيات غير الموسومة، ثم يتدرب النموذج على حل هذه المهمة باستخدام هذه المعطيات، ويتعلم النموذج أثناء التدريب إيجاد تمثيل توليدي شامل (generic representation) لهذه

المعطيات. لاحقاً، يُستخدم هذا التمثيل لحل مسألة جزئية محددة (downstream task) وتكون كلفة التدريب في هذه الحالة منخفضة جداً من ناحية عدد عينات التدريب (Balestriero, et al. 2023). تتكون عملية التعليم الخاضع للإشراف الذاتي من المراحل التالي:

توليد المهمة الأولية من المعطيات غير الموسومة بشكلٍ آلي وفق تقنيات خاصة تناسب طبيعة المعطيات (المهام الأولية المستخدمة في مجال معالجة اللغات الطبيعية تختلف عن مثيلاتها في مجال الرؤية الحاسوبية أو معالجة الكلام).

التدريب الأولي pre-training: يتم تدريب النموذج باستخدام المعطيات غير الموسومة على حَلِّ المهمة الأولية وخلال هذه المرحلة يجري تعلُّم التمثيل التوليدي الشامل.

التوليف الدقيق fine-tune: يتم استخدام التمثيل الشامل الذي تم تعلمه سابقاً كدخل لنموذج أقل تعقيداً من أجل تنفيذ مهمة محددة وتدريب هذا النموذج باستخدام كمية قليلة من المعطيات الموسومة.

على سبيل المثال، في مجال معالجة اللغات الطبيعية للنصوص، يتم توليد المهمة الأولية عبر حجب كلمة من الجملة ومحاولة التنبؤ بها اعتماداً على الكلمات المجاورة، وهذا يشجع النموذج على تعلم اكتشاف الترابط بين الكلمات، ويمكّنه من إيجاد تمثيل عام يعتمد على سياق الجملة وليس فقط على كلماتها، ومن هنا جاءت تسمية التمثيل السياقي (contextual representation). لاحقاً يتم استخدام التمثيل السياقي في العديد من المهام المتعلقة بمعالجة اللغات الطبيعية كالترجمة الآلية والتلخيص الآلي، وغيرها (Balestriero, et al. 2023).

حقق نموذج Bidirectional Encoder Representations from Transformers BERT[4] المطور من قبل شركة Google عام 2018، والذي يستخدم تقنيات التعلم الخاضع للإشراف الذاتي، قفزة نوعية في مجال بناء النماذج اللغوية، واستطاع خلال فترة بسيطة أن يحقق انتشاراً واسعاً وتحول إلى أداةٍ أساسيةٍ استخدمت في العديد من الأبحاث. شجع هذا النجاح الكبير الباحثين –في مجال معالجة الإشارة الكلامية– على محاكاة أسلوب عمل BERT، واشتقاق نماذج تناسب الإشارة الكلامية. ولكن، عند تطبيق تقنية التعلم الخاضع للإشراف الذاتي على الإشارات الكلامية لابد من الانتباه إلى النقاط الأساسية التالية التي تميّز الإشارة الكلامية عن النصوص اللغوية وهي:

إن اللفظ الواحد على دخل الشبكة يتضمن العديد من الوحدات الصوتية (sound units).

لا يوجد معجم يحدد جميع وحدات الصوت خلال مرحلة التدريب الأولي. وبالتالي، لن تكون عملية توليد المهام الأولية بسيطة، وستحتاج إلى تقنيات معقدة نوعاً ما.

تختلف أطوال وحدات الصوت حسب الشخص وطريقة اللفظ، ولا يوجد حدود واضحة للفصل بينها.

لذلك، فإن أي محاولة لتطبيق التعلم الخاضع للإشراف الذاتي على الإشارات الكلامية لابد وأن تلحظ هذه النقاط الثلاث، من أجل الحصول على نتائج جيدة (Hsu, et al. 2021).

3-1-2- آلية نقل التعلم Transfer Learning:

هي إحدى تقنيات تعليم الآلة، تركز على نقل المعارف المكتسبة أثناء حل إحدى المسائل إلى مسألة أخرى تتشابه إلى حدٍ ما مع المسألة الأولى، بحيث يتم اختصار الزمن والموارد اللازمة للتدريب (Hosna, et al. 2022).

هنالك نوعان من تقنيات نقل التعلم، هما التوليف الدقيق fine tune واستخلاص السمات Feature Extraction. في الأول يُستخدم نموذج مدرب مسبقاً(HuBERT في هذه الدراسة) ويضاف نموذج رأسي قليل التعقيد يناسب المهمة الجديدة ويتم تحديث جميع متحولات النموذجين معاً من خلال تدريب النموذج الكلي باستخدام كمية معطيات قليلة. وفي الثاني، يتم البدء باستخدام نموذج مدرب مسبقاً ويضاف نموذج رأسي قليل التعقيد يناسب المهمة الجديدة ولكن يتم تحديث متحولات

[4] BERT: Bidirectional Encoder Representations from Transformers

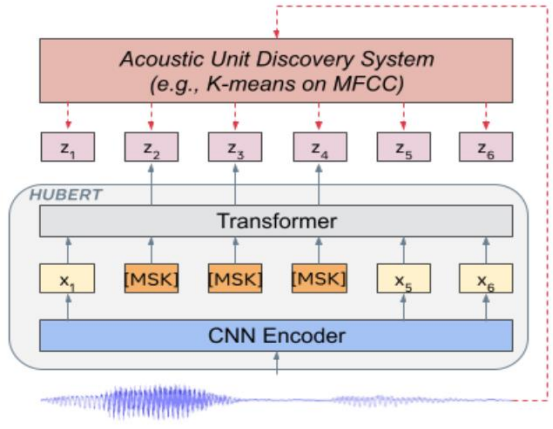
النموذج الرأسي فقط. في هذه الحالة يلعب النموذج المدرب مسبقاً دور مستخرج سمات عميق deep feature extractor للنموذج الرأسي (Zhuang, et al. 2020).

تم في هذا البحث اختبار كلا الطريقتين لمعالجة المسألة المطروحة، ولكن النموذج فشل في التقارب عند محاولة تطبيق نقل التعلم بحالة استخلاص السمات، ولذلك تم اعتماد الطريقة الثانية وهي التوليف الدقيق لكامل النموذج في جميع التجارب المعروضة.

3-1-3- نموذج HuBERT:

هو نموذج لشبكة عصبونية عميقة، يعتمد في بنيته بشكلٍ أساسي على المحولات (Transformers) التي تستخدم تقنية الانتباه متعدد الرؤوس (Multi-Head Attention). تم طرح هذا النموذج عام 2020 من قبل فريق Facebook AI Research (FAIR)، يتوفر منه ثلاث نسخ (versions) تختلف فيما بينها في عدد المحولات (transformers) وبالتالي العدد الكلي لمتحولات النظام (system parameters)، هذه النسخ هي:(Base (95 M Params ، و Large (317 M Params)، و(X-Large (964 M Params (M تشير إلى مليون). تم تدريب هذه النماذج باستخدام 60000 ساعة من الكلام باللغة الإنكليزية من المدونة Libri-light واستُخدم للتدريب 32 وحدة معالجة الرسوميات Graphics Processing Unit (GPU) لتدريب النسخة Base، و128 للنسخة Large، و256 للنسخة X-Large وطرحت هذه النماذج المدربة للاستخدام العام (Hsu, et al. 2021). لاحقاً تم طرح نسخ مدربة من هذه النماذج بلغات أخرى ومنها العربية من قبل بعض المؤسسات. نظراً لضخامة هذه النماذج والقدرات الحسابية الهائلة التي تتطلبها عملية التدريب؛ يلجأ معظم الباحثين إلى الاستعانة بتقنيات نقل التعلم لاستخدام النماذج المدربة مسبقاً وإعادة تدريبها وفق المسألة التي يعالجونها، وقد اتُبع نفس الأسلوب في هذا البحث، حيث تم استخدام نموذجين هما النموذج facebook/hubert-base-ls960 المدرب مسبقاً على اللغة الانكليزية والمقدم من شركة Hugging Face, Inc[5] (von Platen 2021)، والنموذج HuBERT_Large_Arabic المدرب مسبقاً على كلام باللغة العربية باستخدام مدونة تتضمن 2000 ساعة من الكلام باللغة العربية. يبين الشكل (1) المخطط الصندوقي لنموذج HuBERT في طور التدريب.

يتألف النموذج في طور التدريب من أربعة أجزاء هي: مرمز الموجة الصوتية الالتفافي (CNN Encoder) الذي يتألف من سبع طبقات تلفيفية وهو موحَّد في جميع النسخ، مرمز BERT الذي يتكون من عدة كتل من المحولات المتطابقة يختلف عددها حسب النسخة، وطبقة الإسقاط projection layer التي تقوم بنقل التمثيلات المستخرجة من المراحل السابقة إلى فضاء ذي أبعاد أقل مما يجعلها أكثر قابليةً للتخزين والمعالجة، وأخيراً نظام اكتشاف الوحدات الصوتية Acoustic Unit Discovery) (System. وبعد انتهاء عملية التدريب يتم الاستغناء عن كتلة اكتشاف الوحدات الصوتية وطبقة الإسقاط والاستعاضة عنها بنموذج يلائم المهمة المستهدفة، ثم إجراء عملية التوليف الدقيق.



الشكل (1) المخطط الصندوقي لنموذج HuBERT في طور التدريب (Hsu, et al. 2021)

[5] Hugging Face, Inc: an American company that develops tools for building applications using machine learning.

يستخدم HuBERT لاستخراج التمثيل السياقي Contextual Representation للإشارة الكلامية، وهو يعتمد مبدأ التعلم الخاضع للإشراف الذاتي في تدريبه. تقسم عملية التدريب إلى مرحلتين، مرحلة العنقدة (clustering)، ومرحلة التنبؤ المقنع (masked prediction). تهدف مرحلة العنقدة إلى اكتشاف الوحدات الصوتية ليتم استخدامها لاحقاً في اشتقاق أهداف للتدريب من معطيات التدريب ذاتها، وهي الخطوة الأولى في عملية التعلم الخاضع للإشراف الذاتي التي ذُكرت في الفقرة 3-1-1.

إن الإشارة الكلامية هي إشارة مستمرة، ولا يوجد تعريف واضح لوحدات الصوت (على سبيل المثال، في مجال معالجة النصوص تعتبر الحروف هي وحدات الكتابة ومنها يتم تشكيل جميع الكلمات، وتعتبر البكسلات وحدات الصورة ومنها تتكون الصور)، ولهذا السبب تم إضافة المرحلة الأولى في نموذج HuBERT والتي تسمى الوحدة المخفية (Hidden Unit) لتقوم بمهمة اكتشاف الوحدات الصوتية (sound units) عبر استخدام خوارزمية k-mean، وهي خوارزمية عنقدة تعمل وفق مبدأ التعلم غير الخاضع للإشراف وتستخدم لفصل مجموعة من العينات إلى K صف يحددها المستخدم، تعمل وفق مبدأ تكراري من خلال تقسيم العينات إلى K صف أولي؛ يُحسب لكل منها مركز يسمى centroid، ثم يتم تحديد تبعية جميع العينات إلى الصفوف من خلال حساب البعد عن مركز كل عنقود، يتم بعدها تعديل المراكز لكل صف بناءً على حساب الوسطي (mean) لكل عنقود وتتكرر هذه العملية حتى الوصول إلى الدقة المناسبة. بعد الانتهاء من تدريب النموذج على اكتشاف الوحدات الصوتية يتم الانتقال إلى مرحلة تدريبه على اكتشاف التمثيل السياقي من خلال استخدام تقنية التنبؤ المقنَّع (masked prediction)، حيث يتم وضع قناع يخفي جزءاً عشوائياً من معطيات الدخل بنسبة معينة (حوالي 20-50%) ويبقي جزءاً آخر، وهذه التقنية مستخدمة في مجال التعلم الخاضع للإشراف الذاتي من أجل تحريض النظام على التنبؤ بالأجزاء المخفية اعتماداً على الأجزاء غير المخفية، ومن خلال هذه العملية يتعلم كيفية استخراج التمثيل السياقي لإشارة الدخل (Hsu, et al. 2021).

تتحدث الفقرتان الجزئيتان التاليتان عن المدونتين اللتين تم استخدمهما في التدريب والاختبار في هذا البحث.

3-1-4- مدونة Google Speech Command :2.0

تم إطلاق هذه المدونة من قبل Warden بالتعاون مع Google Brain، وتتضمن 105829 مقطعاً كلامياً منفصلاً لـ 35 كلمة باللغة الانكليزية. كل لفظ مسجل على شكل ملف WAV لمدة ثانية واحدة أو أقل. تمت رقمنة العينات (الألفاظ utterances) على 16-bits بقناة أحادية single-channel وبتردد تقطيع 16 kHz. العدد الكلي للمتكلمين 2618 متكلم، والحجم النهائي للمدونة دون ضغط تقريباً 3.8 GB. تقسم القائمة إلى 24 كلمة أساسية و11 كلمة غريبة.

تتضمن الكلمات الأساسية الأعداد من صفر حتى تسعة إضافة إلى 14 كلمة تستخدم كأوامر صوتية للآلة هي: "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", "Go", "Backward", "Forward", "Follow", and "Learn". تتضمن الكلمات الغريبة كلماتٍ مثل "Tree" تم اختيارها لكونها تعطي صوتاً قريباً من بعض الكلمات الأساسية، وبالتالي يمكن استخدامها لاختبار مدى قدرة النظام على التمييز. أما باقي الكلمات فقد اختيرت لتغطي أكبر قدر ممكن من الصوتيمات، وبذلك كانت القائمة النهائية للكلمات الغريبة هي: "Bed", "Bird", "Visual", "Cat", "Dog", "Happy", "House", "Marvin", "Sheila", "Tree", and "Wow". تم جمع العينات آلياً عبر متطوعين من خلال إطلاق واجهة التخاطب البرمجية WebAudioAPI التي تعمل على مستعرضات الأنترنت المشهورة مثل Firefox وChrome والأجهزة الذكية العاملة بنظام Android. تمت مراجعة

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

التسجيلات بشكلٍ آلي كمرحلة أولية تلاها مرحلة مراجعة يدوية نهائية (Warden 2018).

منذ إطلاق هذه المدونة عام 2018 أصبحت بمثابة المرجع الذي يتم من خلاله تقييم أداء نظم تعرُّف الكلمات المفتاحية باللغة الإنكليزية كما هو الحال في **(Seo, Oh and Jung 2021)** و **(Yang, et al. 2021)** و **(Lin, et al. 2020)**.

3-1-5- مدونة Arabic Speech Command:

تم بناء هذه المدونة بشكلٍ مماثل لبنية المدونة google speech command، فهي تتضمن عدداً من التسجيلات الصوتية لكل كلمة مخزنة بصيغة wav ومجمعة ضمن مجلد خاص بكل كلمة ويحمل اسمها. كذلك توفّر المدونة تقسيماً للعينات إلى عينات اختبار test، وعينات تطوير dev، وعينات تدريب train على شكل ملفات(comma-separated values) CSV بحيث تضمن أن جميع التسجيلات الخاصة بمتكلم معين المدونة النهائية من 12000 زوج من هذا القبيل، تضم 40 كلمة مفتاحية. يبلغ طول كل ملف صوتي ثانية واحدة بتردد تقطيع 16 KHz. تتضمن المدونة تسجيلاتٍ لـ 30 شخصاً، سجل كل منهم 10 ملفات لكل كلمة مفتاحية، ليتكون في النهاية 300 ملف صوتي لكل كلمة وفي المجموع (30 * 10 * 40 = 12000 ملف)، والحجم الإجمالي للملفات المسجلة هو 384 MB. تحتوي المدونة أيضاً على العديد من التسجيلات التي تستعمل كضجيج خلفية تم الحصول عليها من مصادر طبيعية مختلفة للضجيج. بحجم إجمالي يبلغ 49 MB. هذه المدونة متاحة بالمجان، واستخدمت في تدريب النموذج المقدم في الورقة البحثية (Salhab and Harmanani 2023).

3-2- طرائق البحث:

تعرض هذه الفقرة النموذج المقترح لحل المسألة، وكيفية تجهيز معطيات التدريب، ومنهجيات اختبار الأداء.

3-2-1- بنية النظام المقترح:

موجودة ضمن إحدى هذه التقسيمات فقط. وبذلك، يمكن اعتبار المدونة بمثابة نسخة معربة من مدونة غوغل، ويستطيع الباحثون المهتمون في مجال تعرُّف الكلمات المفتاحية وتعرف المتكلمين وغيرها من المسائل اختبار نماذجهم على اللغة العربية باستخدام هذه المدونة (Ghandoura, Hjabo and AlDakkak 2021).

لا شك أن مقارنة هذه المدونة مع مدونة غوغل من ناحية عدد العينات لكل كلمة، وعدد المتكلمين، وتنوع المتكلمين غير عادلة، فقد جمعت هذه المدونة بجهد شخصي من الباحث دون الاستعانة بأي شركات، وضمن موارد محدودة جداً، ومع ذلك يمكن اعتبارها أداةً جيدةً وفعالةً لتدريب النماذج التي لا يتطلب تدريبها عدداً كبيراً من العينات، كما هو الحال في هذه المسألة.

تتضمن هذه المدونة قائمة من الأزواج **(x،y)**، حيث **x** هي اللفظ، و **y** هي الوسم المقابل لهذا اللفظ. تتكون

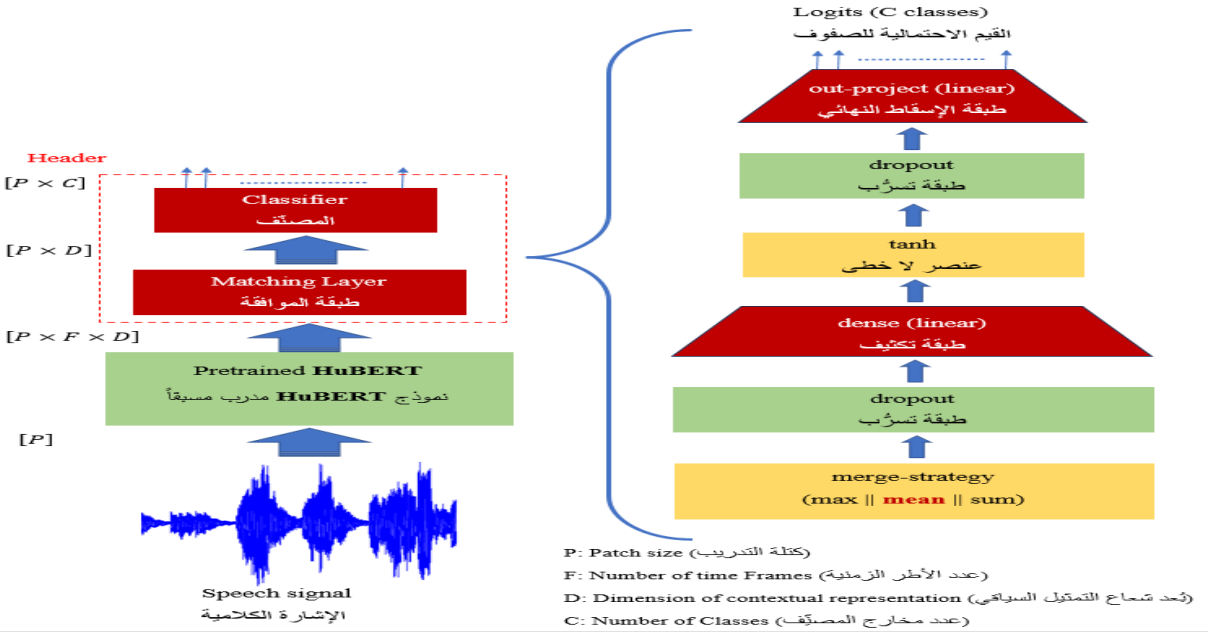
لحل المسألة المطروحة تم اقتراح بنية تتكون من نموذج للتمثيل السياقي مدرب مسبقاً عن طريق التعلم الخاضع للإشراف الذاتي مع نموذجٍ رأسيٍ قليل التعقيد. تم الاستفادة من النتائج المعروضة في الدراسة (Yang, et al. 2021) التي تم فيها اختبار عدة نماذج للتمثيل السياقي مثل wav2vec و wav2vec2.0 و HuBERT وغيرها في مسائل تتعلق بمعالجة الإشارة الكلامية على اللغة الإنكليزية، وبينت هذه الدراسة تفوق نموذج HuBERT على باقي النماذج المعتمدة في تلك الدراسة في مسألة تعرُّف الكلمات المفتاحية من ناحية دقة الكشف وعدد المتحولات، ولهذا السبب تم استخدامه في هذه الدراسة.

تتكون البنية العامة لنظام التعرُّف المقترح كما هو موضح في الشكل (2) من كتلتين أساسيتين: كتلة HuBERT التي تقوم بمهمة استخراج التمثيل السياقي للإشارة الكلامية، والرأس الذي يقوم باستخدام التمثيل السياقي الذي تنتجه الكتلة الأولى وتنفيذ مهمة التعرف على الكلمات المطلوب تعرُّفها وتمييزها عن الكلمات المغايرة. تم شرح بنية الكتلة الأولى في الفقرة 3-1-3، واستخدم في البحث نموذجان مدربان مسبقاً، أحدهما مدرب

على معطيات كلامية (speech) باللغة الإنكليزية واستُخدم في بناء النموذج الانكليزي الذي سُمي KWS_EN، والآخر مدربٌ مسبقاً على معطيات كلامية باللغة العربية واسُتخدم في بناء النموذج العربي الذي سُمي KWS_AR. أما السبب في استخدام نماذج مدربة مسبقاً وليس القيام بتدريب النماذج من الصفر فيعود إلى ضخامة نموذج HuBERT من حيث عدد المتحولات والحاجة إلى موارد حاسوبية ضخمة لا تتوفر لدى الأفراد من أجل تدريبه، ولذلك تم استخدام النسخ المدربة مسبقاً منه المتوفرة عبر الانترنت.

أما الرأس فقد تم بناؤه من قبل الباحث، وهو يتكون من جزئين رئيسيين: طبقة الموافقة Matching layer، والمصنف Classifier وهما موضحان في القسم اليميني من الشكل (2). تتألف طبقة الموافقة من آلية الدمج merge -strategy التي تقوم بتوسيط أشعة التمثيل السياقي الممثلة للكلمة الواحدة، تليها طبقة تسرُّب dropout لتجنب الوقوع في مشكلة التعلم الزائد overfitting أثناء التدريب، ثم يليها طبقة تكثيف dense، وهي طبقة ذات توصيل كامل fully connected layer تقوم بمهمة تعلم تمثيل الروابط المعقدة بين عينات الدخل، يليها عنصر لا خطي tanh لتحسين قدرة الشبكة على نمذجة العلاقات اللاخطية بين عناصر الدخل. تبدأ بعدها كتلة المصنِّف التي تتألف من طبقة تسريب dropout تليها طبقة تكثيف وتسمى عادةً طبقة الإسقاط النهائي out-projection لأنها تقوم بعملية نقل للأشعة الواردة من المرحلة السابقة إلى فضاء جديد تساوي أبعاده عدد مخارج النظام (في مسألة التعرُّف يكون عدد مخارج النظام مساوياً لعدد الكلمات المطلوب تعرُّفها بالإضافة إلى مخرج يمثل جميع الكلمات المغايرة)، ويتم ضبط مخارج هذه الطبقة برمجياً لتتناسب مع عدد الكلمات المطلوب تعرُّفها.

تجدر الإشارة هنا إلى أنه تم اختبار عدة طرق لدمج الأطر في مرحلة الـ Matching، ومنها أخذ القيمة العظمى للشعاع (max)، أو أخذ مجموع قيم عناصر الشعاع (sum) أو أخذ القيمة المتوسطة (mean) حيث أظهرت النتائج أن الأداء الأفضل للنظام كان باستخدام القيمة المتوسطة لذلك تم اعتمادها في هذا البحث.



الشكل (2) بنية النظام المقترح لتعرُّف الكلمات المفتاحية باللغة العربية: نموذج HuBERT مع النموذج الرأسي (يسار الصورة) توضيح لبنية النموذج الرأسي المستخدم (يمين الصورة)

يتميز هذا النوع من شبكات التعلم العميق بأنه يتعامل مع معطيات الدخل الصوتية بصيغتها الأولية، أي أنه لا يحتاج إلى مستخرج سمات مستقل، لأن هذه العملية تعتبر جزءاً من عمل الكتلة الأولى في بنية نموذج HuBERT المستخدم، والتي تم الإشارة إليها بمرمز الموجة الصوتية الالتفافي (CNN Encoder)، حيث تقوم هذه الكتلة بتقسيم إشارة الدخل الكلامية إلى نوافذ زمنية بطول 20 ms (يُفترض خلالها أن الإشارة الكلامية شبه مستقرة لأنها غالباً ضمن نفس الصوتيم)، لتنتج سمات الإشارة ممثلةً على شكل سلسلة زمنية تتكون من F إطار (Frames) لكل مقطع صوتي (يمثل المقطع الصوتي هنا عينة تدريب واحدة من المدونة المستخدمة في التدريب وتمثل في حالتنا كلمة). في طور التدريب يتم إدخال المقاطع الصوتية إلى النموذج على شكل كتل متتالية (patches) تتضمن كل منها P مقطعاً (P=4 في كافة التجارب) ويتم تحديد الخطوة (step = 10 في كافة التجارب) ليتم تحديث أوزان الشبكة بعد كل إدخال لـ P*step عينة تدريب.

تدخل أشعة السمات بعد ذلك إلى القسم الخاص بحساب عينات التمثيل السياقي Contextual Representation الذي يعطي التمثيل السياقي الخاص بكل إطار على شكل شعاع بطول D لنحصل في النهاية على شعاع أبعاده $\mathbf{P \times F \times D}$ لكتلة التدريب الواحدة. يتم اختصار هذا الشعاع من خلال إجراء دمج للأطر الممثلة لكل عينة دخل في مرحلة الـ Matching لنحصل على شعاع أبعاده $\mathbf{P \times D}$ يتضمن التمثيل السياقي لكل عينة ضمن كتلة التدريب P، يستخدم بعدها المصنِّف Classifier هذا التمثيل لتصنيف العينات ضمن كتلة الدخل إلى أحد الصفوف C التي تمثل الكلمات المطلوب تعرُّفها وصفّ الكلمات المغايرة. يتم بعد ذلك تحديث أوزان الشبكة بعد حساب تابع الفقد بناءً على مقارنة نتائج التصنيف مع النتائج الحقيقية المرفقة مع كتلة الدخل، وتتكرر هذه العملية حتى اكتمال عملية التعلُّم والتي سيتم شرح كيفية الاستدلال عليها في الفقرة 3-2-4 (بيئة ومقاييس الاختبار).

في طور الاختبار، يتم إدخال العينات بشكل فردي ويعطي المصنّف قيمة احتمال انتماء هذه العينة إلى كل صف، وبعدها تحدد تبعية الكلمة إلى الصف ذي قيمة الاحتمال الأعظمية.

3-2-2- تجهيز معطيات التدريب:

استُخدمت المدونة Google Speech Command لتنفيذ عملية التوليف الدقيق للنموذج الأجنبي الذي يُرمز له بـ KWS_EN، والمدونة Arabic Speech Command لتنفيذ عملية التوليف الدقيق النموذج العربي الذي يُرمز له بـ KWS_AR، كلا المدونتان مقسمتان إلى جزء للتدريب (training)، وجزء للتطوير (development) وأحياناً يُسمى جزء التحقق (validation)، وآخر للاختبار (test). تُستخدم عينات من قسم التدريب فقط أثناء تدريب النماذج، ويُستخدم الجزء الخاص للتطوير للاختبار وتحديث أوزان الشبكة في طور التدريب، أما عينات الاختبار فتستخدم لتقييم أداء النموذج بعد انتهاء عملية التدريب.

لدراسة أثر محدودية عدد عينات التدريب على أداء النماذج، تم إنشاء آلية لاختيار عدد محدد من عينات التدريب لكل كلمة بشكلٍ عشوائي من ضمن العينات في المدونتين. تقوم الآلية، بعد تحديد العدد المطلوب، بانتقاء عينات عشوائية لكل كلمة من القسم الخاص بالتدريب في المدونة وفق العدد المحدد، دون المساس بعينات التطوير أو الاختبار وتخزين المعطيات المشكَّلة على شكل ملف CSV لاستخدامه في تدريب النماذج.

في المراحل النهائية من الاختبارات تم اختيار قائمة جزئية من الكلمات المتوفرة في المدونة العربية واعتبارها تمثل قائمة الكلمات المطلوب تعرُّفها (12 كلمة من أصل 40)، وفي هذه الحالة تعتبر باقي الكلمات في المدونة كلماتٍ مغايرة، وتعامل جميعها على أنها كلمة واحدة وتصنف على أنها "other"، ويكون عدد الصفوف الكلي على خرج النموذج هو 13. من

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

المؤكد أن قاعدة معطيات التدريب لن تكون متوازنة في هذه الحالة من ناحية العدد لكل كلمة، وسيكون عدد عينات التدريب المصنفة على أنها "other" أكبر بكثير من باقي العينات، وهذا أمرٌ منطقي ويحاكي الواقع، فأثناء التطبيق العملي سيكون عدد الكلمات المطلوب تعرُّفها أقل بكثير من عدد الكلمات المغايرة والتي تمثل جميع مفردات اللغة باستثناء قائمة الكلمات المطلوبة.

3-2-3- منهجية الاختبار:

جرت الاختبارات للبنية المقترحة على ثلاث مراحل:

في المرحلة الأولى تم التركيز على فعالية النموذج المقترح في تعرُّف الكلمات المفتاحية وقدرته على التعلُّم. ولهذه الغاية تم تدريب واختبار النموذج KWS_EN، وشملت قائمة الكلمات المطلوبة جميع الكلمات في المدونة google speech command، واستُخدمت جميع عينات التدريب المتوفرة في المدونة لتدريب النموذج. دُرب النموذج الكلي لـ epoch واحد (يُستخدم مصطلح "epoch" للدلالة على مرور كامل على كافة معطيات التدريب لمرة واحدة). كُررت العملية على النموذج KWS_AR مع المدونة الخاصة به.

في المرحلة الثانية تم دراسة أثر محدودية معطيات التدريب على أداء النموذج العربي KWS_AR فقط، وشملت قائمة الكلمات المطلوب تعرُّفها جميع الكلمات في المدونة العربية (40 كلمة). تم تقييد عدد عينات التدريب لكل كلمة إلى 60، 20، 10، 5 تباعاً (العدد الكلي المتاح هو 180 عينة تدريب لكل كلمة في المدونة Arabic Speech Command)، وتم إجراء التوليف الدقيق للنموذج من نقطة الصفر في كل مرة.

في المرحلة الثالثة تم اختبار أداء النموذج KWS_AR في الحالة المعيارية، حيث يُحدد جزء من الكلمات المتاحة في المدونة من أجل التعرُّف عليه، وتعتبر الكلمات المتبقية على أنها مغايرة "other". تضمنت قائمة الكلمات المطلوب تعرُّفها 12 كلمة، وتم دراسة تقييد عدد عينات التدريب إلى 60، 20، 10، 5 عينة لكل كلمة بهدف مقارنة الأداء مع الحالة السابقة وتحديد الحد الأدنى من عدد العينات المناسب لتحقيق دقة تعرُّف معينة.

3-2-4 بيئة ومقاييس الاختبار TestingMetrics:

تم تطوير البرامج بلغة Python، واستخدمت مكتبة torch لبناء الشبكات العصبونية، واستخدمت مكتبة transformers المطورة من قبل شركة Hugging Face, Inc التي توفر العديد من التوابع التي تسهل وتنظم عمليات التدريب وتعديل الأوزان والاختبار. تم تحديد كتلة التدريب في كافة التجارب patch size =4، وخطوة تعديل الأوزان step = 10، وتمت عمليات التدريب والاختبار باستخدام حاسوب مخصص ذي مواصفات عالية (معالج Intel core i9 من الجيل التاسع، وحدة معالجة الرسوميات NVIDIA GForce RTX 2080 Ti GPU، 64 GB RAM).

لاختبار قدرة النماذج على التعلم أثناء عملية التدريب وتحديد نقطة انتهاء عملية التدريب هنالك عدة طرق ورد ذكرها في المرجع **(Aggarwal 2023)** نذكر منها:

فقد معطيات التحقق (validation loss): حيث يتم حساب الفقد بعد كل خطوة تدريب (step) على معطيات التحقق في قاعدة المعطيات المستخدمة، ويدل تناقص الفقد على أن عملية التعلم تتقدم مع كل خطوة، ويُستدل على انتهاء عملية التدريب أيضاً من خلال هذا المقدار، فعندما تستقر قيمة الفقد أو تعود للتزايد بعد أن كانت في طور التناقص يكون ذلك بمثابة مؤشر على أن النموذج قد وصل حد الإشباع ولن يتعلم أكثر من هذا الحد وبالتالي يجب إيقاف التدريب.

منحنيات التعلُّم Learning Curves: هي منحنيات تعبر عن أداء النظام على معطيات التدريب والتحقق كتابع للزمن، إذا كانت هذه المنحنيات متذبذبة بين الصعود والهبوط فهذا يعني أن النموذج لا يتعلم، كما أن استقرارها بعد فترة من الصعود يعني أن النموذج قد بلغ مرحلة الإشباع وأن عملية التدريب قد بلغت نهايتها ويجب إيقافها.

التحقق من التقارب Checking for convergence: في بعض الحالات، يمكن الاستدلال على تعلُّم النموذج من خلال تقارب متحولاته، ويُستدل على تقاربها من خلال وصولها إلى حالة لا تتغير فيها بشكل ملحوظ مع كل تكرار. وكذلك يشير استقرار المحولات بعد فترة من التدريب إلى أن النموذج قد بلغ مرحلة الاشباع ويمكن إيقاف عملية التدريب.

اعتُمدت طريقة validation loss في هذه الدراسة من أجل التأكد من سير عملية التعلم بالشكل الصحيح وتحديد المرحلة التي يمكن القول فيها أن النموذج قد تعلم وبالتالي يمكن إيقاف التدريب. حيث تم تدريب كل نموذج لعدد من الدورات epochs (يُستخدم المصطلح "epoch" للدلالة على مرور كامل على كافة معطيات التدريب) ويتم زيادة عدد الدورات حتى الوصول إلى النقطة التي يستقر فيها الفقد.

لقياس أداء النماذج المختلفة في مرحلة الاختبار تم استخدام المعاملات[6] التالية:

Macro average: المتوسط الحسابي دون توزين على أساس عدد العينات لكل كلمة.

Weighted average: المتوسط الحسابي مع توزين على أساس عدد العينات لكل كلمة.

الدقة precision: وتعطى بالعلاقة:

$$\boldsymbol{precision} = \frac{\boldsymbol{tp}}{\boldsymbol{tp} + \boldsymbol{fp}} \qquad (\mathbf{1})$$

حيث ($\boldsymbol{tp}$: تمثل عدد مرات الكشف الصحيح true positive) و($\boldsymbol{fp}$: تمثل عدد مرات الإنذار الكاذب false positive) وتعبّر الـprecision عن قدرة النموذج على عدم اعتبار العينات الخاطئة كعينات صحيحة أو ما يسمى معدل الإنذار الكاذب.

الاستعادة recall: وتعطى بالعلاقة:

$$\boldsymbol{recall} = \frac{\boldsymbol{tp}}{\boldsymbol{tp} + \boldsymbol{fn}} \qquad (\mathbf{2})$$

حيث ($\boldsymbol{fn}$: تمثل عدد مرات فقد الهدف false negative) وتعبر الـrecall عن مدى قدرة النموذج على اكتشاف جميع العينات صحيحة.

f1-score (معيار f1): يمكن النظر إليه على أنه المتوسط الموزَّن للدقة precision والاستعادة recall. وفي جميع النتائج المعروضة تم إعطاء نفس الوزن لكلا المعاملين.

الصِّحة Accuracy: تعبر عن العدد الكلي لحالات الكشف الصحيح بالنسبة للعدد الكلي لعينات الاختبار.

تعتبر هذه المعايير من أهم المعاملات المستخدمة في تقييم نظم تعلم الآلة، وهي تفيد في إعطاء فكرة عن مدى قدرة النظام على تحقيق التوازن بين التعلم الزائد (overfitting) والتعلم غير المكتمل (underfitting) **(Huilgol 2023)**. تم التركيز أثناء مناقشة النتائج على معيار الصِّحة Accuracy لكون الهدف الأساسي للمسألة لا يتطلب التركيز على دقة كشف كل كلمة على حدا وإنما دقة الكشف الكلية، وبالتالي اعتبر تحسُّن الصحة دليلاً على تحسن أداء النموذج في جميع المقارنات، كذلك تم إدراج النتائج المحسوبة باستخدام باقي المعايير لإعطاء فكرة عن إداء النظام على مستوى التعرف على كل كلمة على حدا.

4-النتائج والمناقشة Results and Discussion

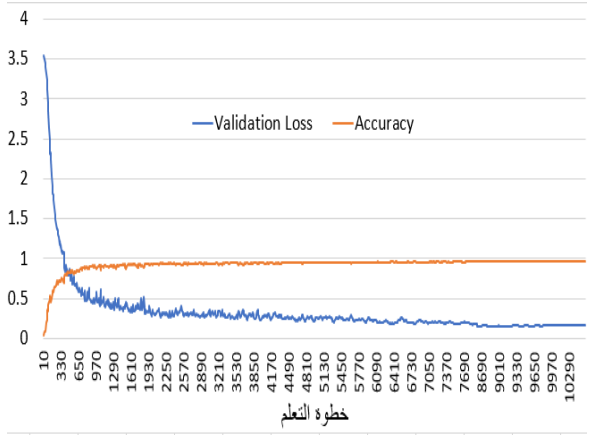
تعرض هذه الفقرة نتائج التجارب التي أُجريت على البنية المقترحة وفق المراحل الثلاث التي ذُكرت في الفقرة 3-2-3.

4-1- نتائج المرحلة الأولى (قدرة النموذج على التعلُّم):

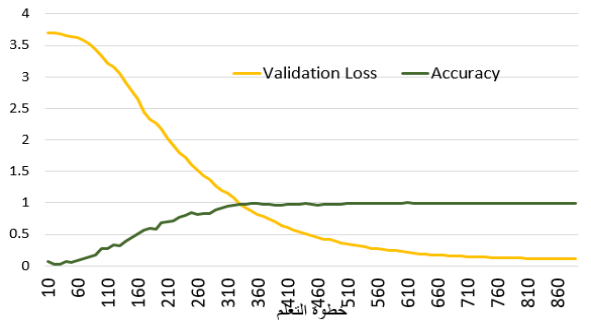
جرى في هذه المرحلة اختبار قدرة النموذجين KWS-EN وKWS_AR على التعلُّم من خلال استخدام المعيار validation loss، وتم رسم تغير قيمة الفقد والصحة بدلالة خطوة التدريب كما هو مبين في الشكلين 3 و4 للنموذج الانكليزي والعربي على الترتيب.

[6] تم حساب جميع المعاملات باستخدام التابع classification_report من المكتبة sklearn.metrics.

يبدو واضحاً من هذين الشكلين أن كلا النموذجين قد بدأ بالتعلم حيث استمر الفقد بالتناقص مع تقدم عملية التدريب، وازدادت الصِحة (Accuracy). نلاحظ أيضاً أن النموذج العربي كان أسرع بالاستقرار حيث وصل بعد حوالي 600 خطوة إلى درجة استقرار مقبولة، في حين احتاج النموذج الانكليزي إلى ما يقارب 10000 خطوة للوصول إلى نفس الدرجة وهذا يعطي فكرة عن مدى الفروقات بين إمكانيات النسخ المختلفة من نموذج HuBERT، حيث استخدمت النسخة Base منه في النموذج الانكليزي، بينما استخدمت النسخة Large في النموذج العربي.



الشكل (3) تطور الفقد والصحة أثناء عملية تدريب النموذج KWS_EN



الشكل (4) تطور الفقد والصحة أثناء عملية تدريب النموذج KWS_AR

يبين الجدولان (1 و2) نتائج الاختبار وفق المعايير التي وردت في الفقرة 3-2-4، حيث تم إيقاف التدريب بعد دورة واحدة. استطاع النموذج العربي الوصول إلى درجة صحة بلغت 99.7%، في حين لم تتجاوز في النموذج الانكليزي 96% بالرغم من كون عدد عينات التدرب في المدونة التي استخدمت لتنفيذ التوليف الدقيق أكبر بكثير من عينات التدريب في المدونة العربية، وهذا يعطي مؤشراً على دور البنية المستخدمة في بناء النموذج، إذ لكل بنية إمكانات لا يمكن أن تتجاوزها مهما ازدادت عدد عينات التدريب.

الجدول (1) نتائج المرحلة الأولى على KWS_EN (جميع عينات التدريب)

	الدقة	الاستعادة	معيار_f1
المتوسط 1*	0.97	0.96	0.96
المتوسط 2**	0.97	0.96	0.97
الصِحة	0.96		

* يُشير المقدار "المتوسط 1" إلى Macro Average

** يُشير المقدار "المتوسط 2" إلى Weighted Average

الجدول (2) نتائج المرحلة الأولى على KWS_AR (جميع عينات التدريب)

	الدقة	الاستعادة	معيار_f1
المتوسط 1	0.9975	0.9975	0.9975
المتوسط 2	0.9976	0.9975	0.9975
الصِحة	0.9975		

4-2- نتائج المرحلة الثانية (محدودية عينات التدريب):

استمرت عملية التدريب في كل تجربة حتى انتهاء عملية التعلم وفقاً لمعيار validation loss الذي تم الحديث عنه في الفقرة 3-2-4. حيث لوحظ أن الوصول إلى نقطة نهاية التعلم تتطلب زيادة عدد دورات التدريب (epochs) كلما قلَّ عدد العينات المستخدمة، ولكن في النهاية حدث التقارب وتعلم النظام. بالنسبة لعدد دورات التدريب التي احتاجها النموذج لكي يتقارب فقد كانت على النحو التالي: (2، 10، 40، 50) دورة من أجل (60، 20، 10، 5) عينة على الترتيب.

يعرض الجدولان (3 و4) نتائج اختبار النموذج وفق إعدادات المرحلة الثانية التي ذُكرت في الفقرة 3-2-3. بلغت صحة الكشف 99.3% عند استخدام 60 عينة تدريب لكل كلمة، وبقيت على هذه القيمة تقريباً عند تقليل عدد العينات حتى 10 عينات، وعند تقليل العدد حتى 5 عينات بدا التراجع في الدقة واضحاً حيث وصلت إلى 98.5%. وفيما يتعلق بالدقة والاستعادة، فقد أعطى النموذج قيماً مرتفعةً لهذين المعياريين مما يدل على قدرته على التعرف والتمييز بين الكلمات

المختلفة. كذلك حدث انخفاض واضح لهاتين القيمتين عند تقليل عدد العينات إلى 5 عينات تدريب للكلمة.

بناءً على هذه النتائج يمكن القول إنه وفق شروط التجربة الحالية يمكن تدريب النموذج باستخدام عشرة عينات لكل كلمة مطلوب تعرُّفها والحصول على صحة كشف حوالي 99.4%، وإن تقليل عدد العينات عن هذا المقدار سيؤدي إلى تراجع الأداء، كما أن زيادته لن تؤثر كثيراً في تحسين الأداء وزيادة الصِّحة.

الجدول (3) نتائج المرحلة الثانية على KWS_AR (60 و20 عينة تدريب لكل كلمة) جميع الكلمات مطلوبة

عدد العينات	60			20		
	الدقة	الاستعادة	معيار_f1	الدقة	الاستعادة	معيار_f1
المتوسط 1	0.9936	0.9933	0.9933	0.9943	0.9942	0.9941
المتوسط 2	0.9935	0.9933	0.9933	0.9943	0.9942	0.9941
الصِّحة	0.9933			0.9941		

الجدول (4) نتائج المرحلة الثانية على KWS_AR (10 و5 عينات تدريب لكل كلمة) جميع الكلمات مطلوبة

عدد العينات	10			5		
	الدقة	الاستعادة	معيار_f1	الدقة	الاستعادة	معيار_f1
المتوسط 1	0.99551	0.99542	0.99539	0.9868	0.9854	0.9851
المتوسط 2	0.99550	0.99541	0.99538	0.9868	0.9854	0.9851
الصِّحة	0.99541			0.9854		

4

3- نتائج المرحلة الثالثة (الاختبار النموذجي):

تمثل هذه المرحلة الاختبار النموذجي لمسألة التعرُّف، وفيها تم اختيار 12 كلمة من كلمات المدونة العربية من أجل تعرُّفها واعتبرت جمع الكلمات المتبقية على أنها مغايرة (other).

احتاج النموذج في هذه المرحلة أيضاً إلى زيادة عدد دورات التدريب epochs عن المرحلة السابقة حتى تحقيق التقارب والاستقرار، حيث تم إجراء (8، 25، 50، 80) دورة من أجل معطيات التدريب منحازة بشكل كبير للعينات المغايرة مما يسبب تأخراً في تقارب النموذج وبالتالي نحتاج إلى زيادة دورات التدريب للوصول إلى نهاية عملية التعلُّم.

بالنسبة لمعيار صحة التعقب، توضح النتائج المعروضة في الجدولين (5 و6) نتائج الاختبار باستخدام معطيات الاختبار في المدونة، حيث يُلاحظ هنا بالمقارنة مع النتائج في الجدولين (3 و4) أن الصحة بدأت بالتراجع بشكلٍ أسرع من المرحلة (60، 20، 10، 5) عينة تدريب لكل كلمة على الترتيب. يمكن أن يعزى هذا السلوك إلى انحياز معطيات التدريب من حيث عدد العينات لصالح العينات المغايرة، ذلك أن عدد الكلمات المغايرة هو 28 كلمة (العدد الكلي للكلمات في المدونة هو 40 كلمة تم اختيار 12 منها لتعرُّفها واعتبرت باقي الكلمات مغايرة)، فإذا كان عدد عينات التدريب لكل كلمة مطلوبة هو N يكون عدد عينات التدريب للكلمات المغايرة 28*N، وهذا يجعل السابقة، حيث بلغت 98.5% عند استخدام 10 عينات لكل كلمة في التدريب في حين كانت 99.5% في المرحلة السابقة من أجل العدد نفسه، وانخفضت عند استخدام 5 عينات فقط إلى 91.99% وهذا يعني أنه في الحالة النموذجية لن تكون 10 عينات تدريب لكل كلمة كافية لتدريب النموذج والحصول على صحة مماثلة للمرحلة السابقة.

الجدول (5) نتائج المرحلة الثالثة على KWS_AR (60 و20 عينة تدريب لكل كلمة) 12 كلمة مطلوبة من أصل 40 كلمة

عدد العينات	60			20		
	الدقة	الاستعادة	معيار_f1	الدقة	الاستعادة	معيار_f1
المتوسط 1	0.99500	0.99858	0.99675	0.99373	0.99977	0.99669
المتوسط 2	0.99837	0.99833	0.99834	0.99799	0.99791	0.99793
الصِحّة	0.99833			0.99791		

الجدول (6) نتائج المرحلة الثالثة على KWS_AR (10 و5 عينات تدريب لكل كلمة) 12 كلمة مطلوبة من أصل 40 كلمة

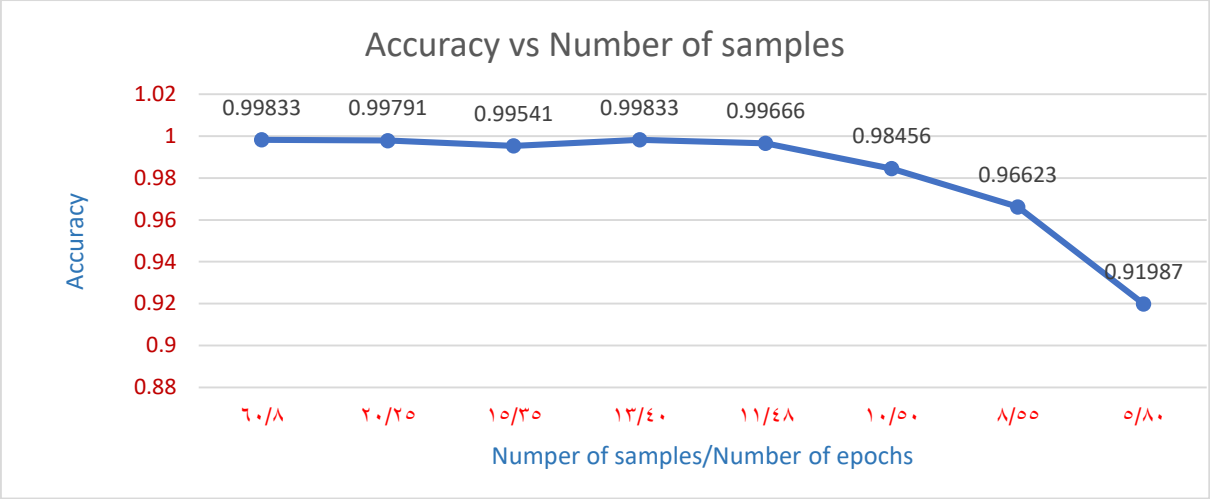
عدد العينات	10			5		
	الدقة	الاستعادة	معيار_f1	الدقة	الاستعادة	معيار_f1
المتوسط 1	0.97890	0.96122	0.96891	0.83430	0.97010	0.88893
المتوسط 2	0.98478	0.98456	0.98428	0.94244	0.91987	0.92527
الصِحّة	0.98456			0.91987		

كذلك يمكن ملاحظة التأثير الكبير لتقليل عدد عينات التدريب على معياري الدقة والاستعادة، فعند استخدام 10 عينات فقط تراجعت الدقة والاستعادة بشكلٍ واضحٍ. وعند استخدام 5 عينات فقط كان التراجع أكبر بكثير حيث وصلت الدقة إلى حوالي 94% والاستعادة إلى حوالي 92% وهذا يعني أن النظام أصبح أقل قدرة على التمييز بين الكلمات المختلفة والتعرف عليها.

بهدف التحديد الدقيق للحد الأدنى من عدد العينات المطلوب والذي يبدأ بعده تناقص الصحة بشكلٍ واضح، تم إجراء المزيد من التجارب عبر إعادة تدريب النموذج وفق نفس إعدادات المرحلة الثالثة باستخدام عدد عينات تدريب (15، 13، 11، 8). تم زيادة عدد دورات التدريب في كل تجربة إلى حين التأكد من اكتمال عملية التعلُّم ووصولها إلى نقطة الإشباع وفق معيار validation loss الذي تم اعتماده في الدراسة، وتم رسم منحني الصحة بدلالة عدد عينات التدريب لكل كلمة فحصلنا على النتائج المبينة في الشكل (5). يُلاحظ أن الصحة بقيت ثابتةً تقريباً بجوار 99.7% حتى وصل عدد العينات إلى 11 عينة (باستثناء انخفاض بسيط عند استخدام 15 عينة حوالي 0.25%)، وبعدها بدأت بالانخفاض بشكلٍ كبيرة (حوالي 1.1%)، واستمر الانخفاض بوتيرة أسرع مع تقليل عدد العينات. قد يكون السبب في تراجع الصِحة قليلاً عند استخدام 15 عينة تدريب هو جودة عينات التدريب المستخدمة في هذه التجربة، إذ أنه وفق الآلية التي اعتمدناها لاختيار عينات التدريب فإن اختيار العينات يتم بشكل عشوائي في كل تجربة. للتأكد من هذه النقطة، قمنا بإعادة نفس التجربة (15 عينة تدريب لكل كلمة) عدة مرات ولكن باستخدام 15 عينة مختلفة في كل مرة، وكانت الدقة مختلفة في كل تجربة ولكن ضمن هامش تغير بسيط (أقل من 0.2%)

بناءً عليه يمكن القول إنه باستخدام 11 عينة تدريب يمكن الوصول إلى صحة تعرُّف حوالي 99.7% وتتناقص الصحة بشكلٍ واضح عند تقليل عدد العينات عن هذا الحد.

تعتبر هذه الحالة أصعب من الحالة السابقة، ذلك أن اكتشاف وتمييز الكلمات المغايرة يزيد من تعقيد مسألة الفصل التي يتم تدريب النموذج الرأسي عليها في مرحلة التدريب، حيث يُطلب من هذا النموذج إيجاد السطح الفاصل (hyper plane) بين كل من الكلمات المطلوبة منفردة وجميع الكلمات المغايرة مجتمعةً– التي بدورها قد تكون متقاربة كثيراً من بعض الكلمات المطلوبة في فضاء الفصل– وهذا يجعل مهمة إيجاد السطح الفاصل أصعب وتتطلب مزيداً من التدريب للوصول إلى نتائج جيدة. ولكن بشكلٍ عام تبقى النتائج مقبولة طالما بقي عدد العينات المستخدمة في التدريب أكبر من 10 عينات لكل كلمة.



الشكل (5) صحة التعرُّف كتابع لعدد عينات التدريب (Number of samples) لكل كلمة مع ذكر عدد دورات التدريب (Number of epochs) التي احتاجها النموذج للتقارب

بالمقارنة مع النتائج المعروضة في الدراسات المرجعة التي تم الاطلاع عليها نجد أن البنية المقترحة تفوقت على النموذج AraSpot المقدم في (Salhab and Harmanani 2023) لتعرُّف الكلمات المفتاحية باللغة العربية والتي تستخدم نفس المدونة التي استُخدمت في هذا البحث في التدريب والاختبار من حيث صحة الكشف (99.7% مقابل 99.58%)، كما أنها تفوقت على البنية WAV2KWS المقدمة في (Seo, Oh and Jung 2021) لتعرُّف الكلمات المفتاحية باللغة الكورية من خلال القدرة على تحقيق دقة كشف 98.5% باستخدام 10 عينات تدريب لكل كلمة مقابل 95% حيث يلخص الجدول (7) هذه المقارنة

الجدول (7) مقارنة أداء نموذج تعرُّف الكمات المفتاحية باللغة العربية KWS_AR مع بعض النماذج الأخرى

النموذج	المرجع	البنية	اللغة	المدونة	عدد عينات التدريب	الصحة %
WAV2KWS	(Seo, Oh and Jung 2021)	WAV2VEC2.0 + Header	الكورية	خاصة	10 لكل كلمة	95
AraSpot	(Salhab and Harmanani 2023)	ConformerGRU	العربية	Arabic Speech Command	180+ لكل كلمة	99.58
KWS_AR	الدراسة الحالية	HuBERT_Large + Header	العربية	Arabic Speech Command	10 لكل كلمة	98.5
KWS_AR	الدراسة الحالية	HuBERT_Large + Header	العربية	Arabic Speech Command	11 لكل كلمة	99.7
CRNN_KWS	(Arık, et al. 2017)	مزج بين شبكة التفافية مع شبكة عودية	الانكليزية	تعرُّف كلمة مفردة (“TalkType”)	16000	97.71
ResNet_KWS	(Tang and Lin 2018)	Residual Network	الانكليزية	Google Speech Command	4000-1500 لكل كلمة	95.8

5- الاستنتاجات والآفاق المستقبلية:

قدمت هذه الورقة البحثية دراسةً مفصلةً عن مسألة تعرُّف الكلمات المفتاحية، وتم اقتراح بنية خاصة لتعرُّف الكلمات المفتاحية باللغة العربية تستخدم نموذجاً للتمثيل السياقي بهدف الوصول إلى نظام قادر على تحقيق وظيفة التعرُّف بدقة جيدة وباستخدام عدد قليل من عينات التدريب. نجحت البنية

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

المقترحة في تحقيق دقة تعرُّف بلغت 99.7% باستخدام 11 عينة تدريب فقط لكل كلمة. وعند تجربة أدائها على اللغة الانكليزية كانت النتائج جيدة أيضاً (تجاوزت الدقة 96% رغم أن عملية التدريب لم تستكمل حتى النهاية) وبالتالي يمكن القول إنها مستقلة عن اللغة كما أنها تشكل خطوة إلى الأمام في حل مشكلة تغيير قائمة الكلمات المطلوب تعرُّفها، إذ أن جمع 11 عينة لكل كلمة يُعتبر أمراً قابلاً للتحقيق عملياً إذا ما قورن بجمع 4000 عينة لكل الذي تحتاجه بعض الشبكات التقليدية للتقارب وتحقيق الدقة المطلوبة. يمكن تطوير هذا العمل مستقبلاً من خلال التركيز على النقاط التالية:

1. لضغط أكثر باتجاه تقليل عدد العينات اللازمة للتدريب عن الحد الذي توصلت إليه هذه الدراسة وذلك من خلال الاستفادة من الأفكار الجديدة والآفاق التي يفتحها مجال تعلم التعلم Meta-learning لتقليل عدد عينات التدريب والاكتفاء بعينة واحدة فقط إذا أمكن وهو ما يسمى بالتعلم من عينة واحدة One-shot learning.

2. لعمل على تطوير النموذج بحيث يكون قادراً على تعرُّف الكلمات ضمن سياق الكلام المستمر وعدم الاكتفاء بتعرُّف كلمات منفصلة وبالتالي الانتقال إلى المسألة الأعم وهي تعقب الكلمات المفتاحية.

3. حاولة تقليل التعقيد الحسابي للنموذج النهائي، بحيث يتم الوصول إلى بنية تحقق إضافةً إلى المزايا السابقة إمكانية تشغيلها على حواسب صِغريّة Single Board Computer (SBC) ذات قدرات حسابية متواضعة.

4. طوير البنية لتكون قادرة على العمل في الزمن الحقيقي، وذلك من أجل استخدامها في تطبيقات المراقبة الآنية للكلام.

التمويل: هذا البحث ممول من جامعة دمشق وفق رقم التمويل(501100020595).

References:

1. Aggarwal, Charu. 2023. Aggarwal, C. (2023). Deep Learning: Principles and Training Algorithms. Springer International Publishing.
2. Almutairi, Zaynab, and Hebah Elgibreen. 2023. "Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning." IEEE Access 72134-72147.
3. Arık, Sercan Ö., Markus Kliegl, Rewon Child, Joel Hestness, Andrew Gibiansky, Chris Fougner, Ryan Prenger, and Adam Coates. 2017. Convolutional Recurrent Neural Networks for Small-Footprint Keyword Spotting. arXiv preprint.
4. Awaid, Mostafa, Sahar Fawzi, and Ahmed Kandil. 2014. " Audio Search Based on Keyword Spotting in Arabic Language." International Journal of Advanced Computer Science and Applications 128-133.
5. Balestriero, Randall, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldsteiny, Florian Bordes , et al. 2023. A Cookbook of Self-Supervised Learning. arXiv preprint arXiv:2304.12210.
6. Brik, Youcef, Youcef Chibani, Bilal Hadjadji, and Et-Tahir Zemouri. 2014. "Keyword-guided Arabic word spotting in ancient document images using Curvelet descriptors." 2014 International Conference on Multimedia Computing and Systems (ICMCS). Marrakech, Morocco: IEEE. 57-61.
7. Cheikh Rouhou, Ahmed, Yousri Kessentini, and Slim Kanoun. 2019. "Hybrid HMM/DNN System for Arabic Handwriting Keyword Spotting." Image Analysis and Recognition 216–227.
8. Chen, Guoguo, Carolina Parada, and Georg Heigold. 2014. "Small-Footprint Keyword Spotting Using Deep Neural Networks." ICASSP. Florence, Itaaly.
9. Coucke, Alice, Mohammed Chlieh, Thibault Gisselbrecht, David Leroy, Mathieu Poumeyrol, and Thibaut Lavril. 2019. "Efficient Keyword Spotting Using Dilated Convolutions And Gatinga." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom. 6351-6355.
10. Fathallah, Abir, Mounim El-Yacoubi, and Najoua Ben Amara. 2023. "Transfer Learning for Word Spotting in Historical Arabic Documents Based Triplet-CNN." 18th International Conference on Computer Vision Theory and Applications. SCITEPRESS-Science and Technology. 520-527.
11. Ghandoura, Abdulkader, Farouk Hjabo, and Oumayma AlDakkak. 2021. "Building and benchmarking an Arabic Speech Commands dataset for small-footprint keyword spotting." Engineering Applications of Artificial Intelligence (102).
12. Hosna, Asmau, Ethel Merry, Jigmey Gyalmo, Zulfikar Alom, Zeyar Aung, and Mohammad Abdul Azim. 2022. "Transfer learning: a friendly introduction." (Journal of Big Data) 9.
13. Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 3451-3460.

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

14. Huilgol, Purva. 2023. Precision and Recall | Essential Metrics for Machine Learning. 12 21. Accessed 1 30, 2024. ://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/
15. Keshet, Joseph, and Samy Bengio. 2009. Speech and Speaker Recognition: Large Margin and Kernel Methods. John Wiley & Sons.
16. Lin, James, Kevin Kilgour, Dominik Roblek, and Matthew Sharifi. 2020. "Training Keyword Spotters With Limited And Synthesized Speech Data." 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. 7474-7478.
17. Mohamed, Omar, and Salah A. Aly. 2021. "Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset." arXiv:2110.04425.
18. Sainath, Tara N., and Carolina Parada. 2015. "Convolutional Neural Networks for Small-footprint Keyword Spotting." International Speech Communication Association. Dresden, Germany.
19. Salhab, Mahmoud, and Haidar Harmanani. 2023. "AraSpot: Arabic Spoken Command Spotting." arXiv preprint arXiv:2303.16621 .
20. Seo, Deokjin, Heung-Seon Oh, and Yuchul Jung. 2021. "Wav2KWS: Transfer Learning From Speech Representations for Keyword Spotting." IEEE Access 9: 80682-80691.
21. Tabibian, Shima, Ahmad Akbari, and Babak Nasersharif. 2018. Information Sciences 157-171.
22. Tabibian, Shima, Akram Shokri, Ahmad Akbari, and Babak Nasersharif. 2011. "Performance evaluation for an HMM-based keyword spotter and a Large-margin based one in noisy environments." Procedia Computer Science 1018-1022.
23. Tang, Raphael, and Jimmy Lin. 2018. "Deep Residual Learning For Small-Footprint Keyword Spotting." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 5484-5488.
24. von Platen, Patrick. 2021. Hugging Face's logo Hugging Face. 10 7. Accessed 6 5, 2023. https://huggingface.co/facebook/hubert-base-ls960
25. Waheed, Abdul, Bashar Talafha, Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. "VoxArabica: A Robust Dialect-Aware Arabic Speech Recognition System." arXiv preprint arXiv:2310.11069 (Association for Computational Linguistics) 441--449.
26. Warden, Pete. 2018. "Speech commands: A dataset for limited-vocabulary speech recognition." arXiv preprint, arXiv:1804.03209.
27. Yang, Shu wen, Po Han Chi, Yung Sung Chuang, Cheng I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, et al. 2021. "Superb: Speech processing universal performance benchmark." arXiv preprint.
28. Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, and Hui Xio. 2020. "A Comprehensive Survey on Transfer Learning." Proceedings of the IEEE 43-76.