Journal of Education and Science (ISSN 1812-125X), Vol: 31, No: 01, 2022 (93-110)

Intelligence System for Multi-Language Recognition

Fawziya M. Ramo[1*], Mohammed N. Al-Hamdani[2]

[1*,2]Department of Computer Sciences, College of Computer Sciences and Mathematics, University of Mosul, Mosul, IRAQ

Email: [1*]fawziyaramo@uomosul.edu.iq, [2]mohammed.csp61@student.uomosul.edu.iq

(Received April 07, 2021; Accepted July 16, 2021; Available online March 01, 2022)

DOI: 10.33899/edusj.2021.129868.1156, © 2022, College of Education for Pure Science, University of Mosul.

ABSTRACT

Language classification systems are used to classify spoken language from a particular phoneme sample and are usually the first step of many spoken language processing tasks, such as automatic speech recognition (ASR) systems Without automatic language detection, spoken speech cannot be properly analyzed and grammar rules cannot be applied, causing failures Subsequent speech recognition steps. We propose a language classification system that solves the problem in the image field, rather than the sound field. This research identified and implemented several low-level features using Mel Frequency Cepstral Coefficients, which extract traits from speech files of four languages (Arabic, English, French, Kurdish) from the database (M2L_Dataset) as the data source used in this research.

A Convolutional Neuron Network is used to operate on spectrogram images of the available audio snippets. In extensive experiments, we showed that our model is applicable to a range of noisy scenarios and can easily be extended to previously unknown languages, while maintaining classification accuracy. We released our own code and extensive training package for language classification systems for the community.

CNN algorithm was applied in this research to classify and the result was perfect, as the classification accuracy reached 97% between two languages if the sample length was only one second, but if the sample length was two seconds, the classification accuracy reached 98%. While the classification among three languages, the classification accuracy reached 95% if the sample length was only one second, but if the sample length was two seconds, the classification accuracy reached 96%.

Keywords: Language Classification, Mel-frequency cepstral coefficients (MFCC), Convolutional Neural Networks, Deep Learning.

نظام ذكائي للتعرف على لغات متعددة

فوزية محمود رمو[1*]، محمد نايف الحمداني[2]

[2،*1] قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

الخلاصة

تستعمل أنظمة تصنيف اللغة لتصنيف اللغة المنطوقة من عينة صوتية معينة وهي عادةً الخطوة الأولى للعديد من مهام معالجة اللغة المنطوقة، مثل أنظمة التعرف التلقائي على الكلام ومن دون الاكتشاف التلقائي للغة لا يمكن تحليل الكلام المنطوق بشكل صحيح ولا

يمكن تطبيق القواعد النحوية، مما يتسبب في فشل خطوات التعرف على الكلام اللاحقة. نقترح نظام تصنيف اللغة الذي يحل المشكلة في مجال الصورة، بدلاً من مجال الصوت. حدد هذا البحث ونفذ العديد من الميزات منخفضة المستوى باستخدام معاملات درجة النغم (Mel Frequency Cepstral Coefficients)، والتي تستخلص الصفات من ملفات الكلام لأربع لغات (العربية، الإنجليزية، الفرنسية، الكردية) من قاعدة البيانات (M2L_Dataset) هي مصدر البيانات المستخدمة في هذا البحث.

تُستخدم الشبكة العصبية الالتفافية (Convolutional Neuron Network) بحيث تعمل على صور المخطط الطيفي للمقتطفات الصوتية المتوفرة. أظهرنا في تجارب مكثفة أن نموذجنا قابل للتطبيق على مجموعة من السيناريوهات الصاخبة ويمكن بسهولة توسيعه ليشمل لغات غير معروفة سابقًا، مع الحفاظ على دقة التصنيف. أصدرنا الكود الخاص بنا ومجموعة تدريب واسعة النطاق لأنظمة تصنيف اللغة للمجتمع. تم تطبيق خوارزمية الشبكات العصبية الالتفافية (CNN) في هذا البحث للتصنيف وكانت النتيجة مثالية، إذ بلغت دقة التصنيف 97% بين لغتين إذا كان طول العينة ثانية واحدة فقط، اما إذا كان طول العينة ثانيتين فقد بلغت دقة التصنيف 98%. في حين التصنيف بين ثلاث لغات فقد بلغت دقة التصنيف 95% إذا كان طول العينة ثانية واحدة فقط، اما إذا كان طول العينة ثانيتين فقد بلغت دقة التصنيف 96%.

الكلمات المفتاحية: تصنيف اللغة، معاملات درجة النغم (MFCC)، الشبكات العصبية الالتفافية، التعلم العميق.

1- المقدمة

هناك عدة آلاف من اللغات التي يتم التحدث بها في جميع أنحاء العالم. على الرغم من أن معظمها لا يستخدم إلا من قبل مجموعة صغيرة من الناس، تشير الأبحاث إلى أن أكثر 30 لغة يتحدث بها 40 مليون شخص على الأقل. أي برنامج يمكن للبشر التواصل والتفاعل معه لفظيًا يجب أن يكون قادرًا على فهم أنماط الكلام البشري بكل تفاصيله من النغمة والصوتيات والتردد والكلمات وتركيبات الجمل. من شأن الوظيفة المناسبة لتصنيف اللغات أن تسهل بشكل كبير عند تفسير الكلام البشري. ومن شأنه أيضًا التأكد من أن الكلمات التي يمكن أن يكون لها معان مختلفة في لغات مختلفة يتم تخصيص المعنى الصحيح لها اعتمادًا على اللغة التي يعتقد الكمبيوتر أنه يتم التحدث بها [1].

يستطيع الانسان من تمييز لغة المتكلم في غضون ثواني من سماعها ويمكنه معرفة إذا كانت اللغة لغة الأُم أم لا حيث يعد البشر أكثر أنظمة تحديد اللغة دقة في العالم اليوم [2].

تعتبر الشبكات العصبية الالتفافية (CNN) من أفضل الخوارزميات التي تستخدم لتحديد اللغة [3,4]، وذلك باستخدام المخططات الطيفية للصوت (Language Identification For Audio Spectrums)، حيث تُستخدم المخططات الطيفية لإشارات الصوت الخام كمدخل إلى الشبكة العصبية الالتفافية (CNN) لاستخدامها في تعريف لغة المتكلم، تم اختيار الشبكة العصبية الالتفافية في هذا البحث لأنها تتمتع بالمزايا التالية [5]:

- إحدى مزايا هذه الطريقة في أنها تتطلب القليل جدًا من المعالجة المسبقة. في الواقع، تغذى البيانات الصوتية الأولية فقط في الشبكة العصبية، مع تشكيل مخططات طيفية حيث يتم تغذية كل دفعة في الشبكة.
- ميزة أخرى هي أن التقنية يمكن أن تستخدم مقاطع صوتية قصيرة (حوالي ثانية او ثانيتين) من أجل التصنيف الفعال، وهو أمر ضروري للمساعدين الصوتيين الذين يحتاجون إلى تحديد اللغة بمجرد أن يبدأ المتحدث في التحدث.

- تُقارن الشبكات العصبية الالتفافية (CNN)الصورة جزءاً جزءاً. لذا فإنها ترى التشابه بشكل أفضل من مخططات مطابقة الصورة بأكملها. [5]

على حد علمنا، لم يقم أي باحث سابقاً باستخدام اللغة العربية واللغة الكردية للتصنيف بالإضافة الى ان الدقة التي تم الحصول عليها عند التمييز بين اللغة العربية والانكليزية كانت (98%) في حين حصل النموذج المقترح على دقة تصنيف (96%) عند التصنيف بين ثلاث لغات وهي اللغة العربية والكردية والانكليزية.

هيكلية هذا البحث كما يلي: الفقرة الاولى هي المقدمة، الفقرة الثانية بعض الأعمال ذات الصلة في طرق تصنيف اللغات. الفقرة الثالثة توضح عملية استخلاص الميزات باستخدام معاملات درجة النغم (MFCC). الفقرة الرابعة تشرح بشكل مبسط الشبكة العصبية الالتفافية (CNN) الخوارزمية المستخدمة في النموذج المقترح. الفقرة الخامسة تفاصيل ثلاث لغات (العربية، والكردية، والإنجليزية)، مع اختلاف اللهجات والأجناس التي تستخدم كمجموعات بيانات. الفقرة السادسة توضح كيفية تنظيم بيانات الإدخال لخوارزمية الشبكة العصبية الالتفافية (CNN)، الفقرة السابعة تشرح دوال التنشيط المستخدمة في النموذج المقترح، والفقرة الثامنة توضح نتائج النموذج المقترح، وأخيرًا، الفقرة التاسعة للاستنتاجات.

2- الاعمال ذات الصلة

تناولت البحوث في السنوات الأخيرة عملية تمييز وتصنيف اللغات حول العالم وأجريت دراسات كثيرة حول الموضوع ولخص ما توصل اليه الباحثون السابقون كالاتي:

في عام 2015 قام الباحثان Yaakov HaCohen-Kerner and Ruben Hagege باستخدام تقنيات التعلم الآلي لتصنيف ملفات الكلام لـ7 لغات مختلفة (الفرنسية والفارسية واليابانية والكورية والصينية والتأميل والفيتنامية) بناءً على مجموعات ميزات RASTA ومجموعة ميزات الطيف. تقارن هذه المنهجية بين ست طرق مختلفة لتعلم الآلة وهي (J48 وRandom Forest وMultiBoostab وBayesNet وLogistic Regression وSequential Minimal Optimization)، وحققت خوارزمية الغابة العشوائية (Random Forest) أفضل طريقة التعلم الآلي، إذ حققت نتائج بدقة عالية نسبيًا بنسبة 89.18٪ و81.85٪ و80.33٪ لتجارب التصنيف الآتية: لغتين وخمسة لغات وسبعة لغات على التوالي [6].

في عام 2019 قام الباحثان Shauna Revay and Matthew Teschke باستخدام تقنية موصوفة بالتعرف على اللغة للمخططات الطيفية الصوتية(LIFAS) ، وهي مخططات طيفية لإشارات الصوت الخام كمدخلات إلى شبكة عصبية التفافية (CNN)لاستخدامها في تحديد اللغة. ومن فوائد هذه العملية أنها تتطلب الحد الأدنى من المعالجة المسبقة. في الواقع، يتم إدخال الإشارات الصوتية الأولية فقط في الشبكة العصبية الالتفافية (CNN)، مع إنشاء مخططات الطيف إذ يتم إدخال كل دفعة إلى الشبكة أثناء التدريب. والطريقة المقترحة يمكن أن تستخدم مقاطع صوتية قصيرة (حوالي 4 ثوان) من أجل التصنيف الفعال، توصل الباحثان الى دقة تصنيف بين لغتين تصل إلى 97٪، في حين كانت دقة التصنيف بين ست لغات وهي (إنكليزية والألمانية والإيطالية والفرنسية والاسبانية والروسية) تصل الى 89٪ [5].

في عام 2020 قامت الباحثة Alexandra Draghici وآخرون بإجراء طريقتان مختلفان من اجل تصنيف اللغات وذلك باستخدام خوارزميات الشبكات العصبية الالتفافية (CNN) والشبكات العصبية الالتفافية التكرارية(CRNN) ، وقام الباحثون باستخراج الخصائص باستخدام ميزات طيف الميل (Mel Spectrograms) اما مجموعة البيانات المستخدمة كانت مستودع الاخبار الاوربية (European News Repository)، توضح الطريقة الاولى مقارنة في التصنيف خوارزميات ال (CNN) وال

(CRNN) باستخدام اربع لغات بينما الطريقة الثانية استخدام خوارزمية الـ(CNN) من اجل التصنيف بين ست لغات وكما موضح ادناه:

الطريقة الأولى: التصنيف المكون من أربع لغات (الإنكليزية والألمانية والفرنسية والاسبانية) يحقق نموذج CNN معدلات تصنيف جيدة جدًا حيث بلغت دقة التدريب 97% بينما دقة الاختبار 82%. ومن خلال إضافة طبقة LSTM ثنائية الاتجاه المتكررة في نموذج CRNN، يمكن ملاحظة تحسن طفيف فقط بمقدار 0.01 لكل من مقاييس الدقة والاختبار، مما يشير إلى عدم وجود فائدة فعلية لنماذج CRNN لهذه المهمة.

الطريقة الثانية: التصنيف المكون من ست لغات (الإنكليزية والألمانية والفرنسية والاسبانية والإيطالية واليونانية) قام الباحثون بتقييم نموذج CNN في مستودع الأخبار الأوروبية لسيناريو اللغات الست. على الرغم من التعقيد المتزايد للمهمة، يحقق النموذج قيم دقة جيدة تبلغ 87% للتدريب و71% للاختبار، مما يؤكد أن نماذج CNN مناسبة للمهمة قيد البحث [7].

في عام 2020 قامت الباحثة Anna Avenberg بتقديم نموذج من نوع الذاكرة طويلة–قصيرة المدى (Long–Short Term Memory) من اجل تصنيف اللغات عن طريق النصوص، اللغات التي تم تصنيفها هي (الكرواتية والإنكليزية والألمانية والبولندية والبرتغالية والروسية والصربية والاسبانية والسويدية). يعد تويتر (Twitter) مكانًا جيدًا للعثور على العديد من كتابات الأشخاص المختلفين، تم بناء نموذج التعلم الآلي للذاكرة طويلة–قصيرة المدى (LSTM) لبيانات تغريدات تويتر ومقارنة النتائج مع نموذج (Fast Text) الخاص بـ Facebook. قامت الباحثة باستخلاص الصفات باستخدام طريقة حقيبة الكلمات (The Bag of Words) وطريقة (N–grams) .تظهر النتائج كيف وصل نموذج الذاكرة طويلة–قصيرة المدى (LSTM) إلى دقة حوالي 95٪ ونموذج (Fast Text) المستخدم للمقارنة وصل إلى دقة 97٪. كان أداء تصنيف نموذج الذاكرة طويلة–قصيرة المدى (LSTM) ضعيفًا نسبيًا في الحالات التي كانت فيها اللغات متشابهة، مثل الكرواتية والصربية. وكانت نتائج نموذج LSTM ونموذج (Fast Text) إلى دقة 94٪ [8].

في عام 2021 قام الباحث Herman Groenbroek بتقديم منهجية اسماها (VGGish) واستخدامها للتصنيف بين ست لغات (الإنكليزية، الهولندية، الألمانية، الفرنسية، الاسبانية، البرتغالية) بالاعتماد على مجموعة بيانات موسيقية جديدة سُميت (L5K 6 Music Corpus). يتم الحصول على مجموعة بيانات الجزء الصوتي عن طريق أخذ أجزاء صوتية مدتها 3 ثوانٍ من مجموعة الأغاني الموسيقية (L5K6) ومقارنة المنهجية المقترحة مع الشبكة العصبية العميقة (Deep Neuron Network). استخراج الصفات تم باستخدام معاملات درجة النغم (Mel–Frequency Cepstral Coefficients).

تشير النتائج في هذه الرسالة إلى أن مهمة تمييز اللغة للأغاني الموسيقية في الشبكة العصبية العميقة (DNN) حصلت على دقة تدريب تصل إلى 35٪ لست لغات. بينما حصل النظام المقترح (VGGish) الى دقة تدريب تصل إلى 41٪ في نفس مجموعة البيانات ذات الست فئات. عند استخدام هذه الأنظمة على بيانات الاختبار كانت دقة الشبكة العصبية العميقة (DNN) 18.1٪، بينما كانت دقة نظام (VGGish) 35.2٪ [9].

3– استخلاص الصفات(Features Extraction)

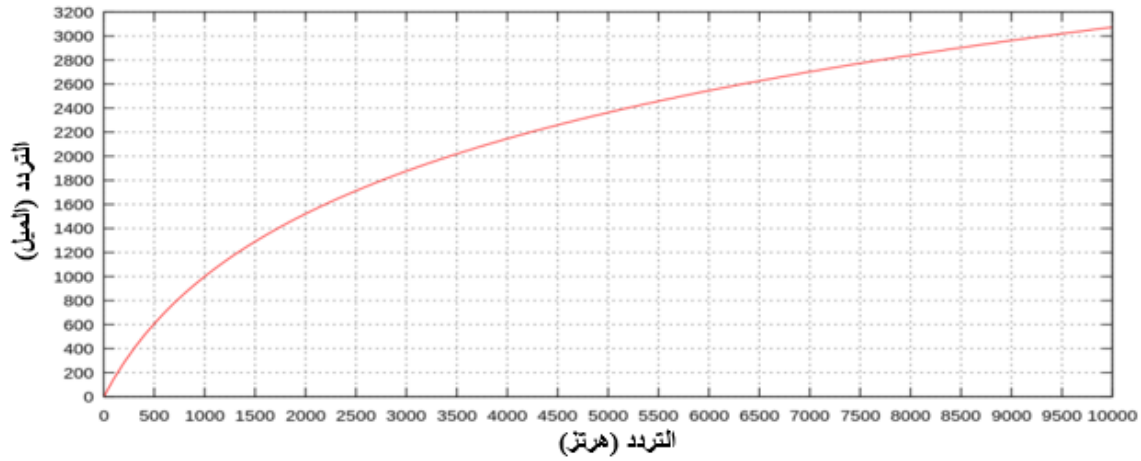
لتصنيف إشارة واردة، تستخلص بعض الصفات منها. يتم تمثيل مجموعة من ميزات D المستخرجة من تحليل أو نافذة نسيج على شكل متجه D الأبعاد C=[C1,C2,……,CD]T يسمى متجه الميزات. النقطة الأساسية هي أن الميزات المختارة يجب أن تحتوي على معلومات قيمة تسمح بالتمييز بشكل صحيح بين الفئات المدروسة. بمعنى آخر، يجب أن تقيس الميزات خصائص الإشارة التي تميل إلى تقديم قيم يمكن تمييزها بين فئات الصوت المختلفة [10].

3-1 معاملات درجة النغم ((Mel Frequency Cepstral Coefficients (MFCC))

إن معاملات درجة النغم هي مجموعة من الميزات الصوتية الطيفية المستخدمة في أنظمة التعرف على اللغات. يتم تمثيل درجة النغم بقائمة من المعاملات تسمى معاملات درجة النغم (MFCC) [11].

القوقعة عبارة عن جهاز لمعالجة الصوت في الأذن يفسر ترددات الصوت. تم تصميم معاملات درجة النغم لتقليد وظائف القوقعة. يتم ذلك عن طريق حساب ترددات الصوت المختلفة أولاً. يتم ذلك لأن جدران القوقعة مبطنة بشعر صغير يهتز اعتمادًا على الترددات في الصوت. تواجه قوقعة الأذن صعوبة في التمييز بين الأصوات التي لها اختلافات صغيرة في التردد [12].

يوفر النظام السمعي المحيطي البشري الأساس لمعاملات درجة النغم (MFCC). لا يدرك البشر محتوى تردد الأصوات لإشارات الكلام على مقياس خطي. وبالتالي، تُقوم درجة الصوت الذاتية على مقياس يسمى مقياس ميل (Mel Scale) لكل نغمة بتردد فعلي يُقاس بالهرتز. يستخدم مقياس ميل تباعد تردد لوغاريتمي أقل من 1000 هرتز وتباعد تردد خطي يتجاوز 1 كيلو هرتز. يتم تحديد نغمة 1 كيلو هرتز، 40 ديسيبل فوق عتبة السمع الحسي، على أنها 1000 ميل كنقطة مرجعية [13]



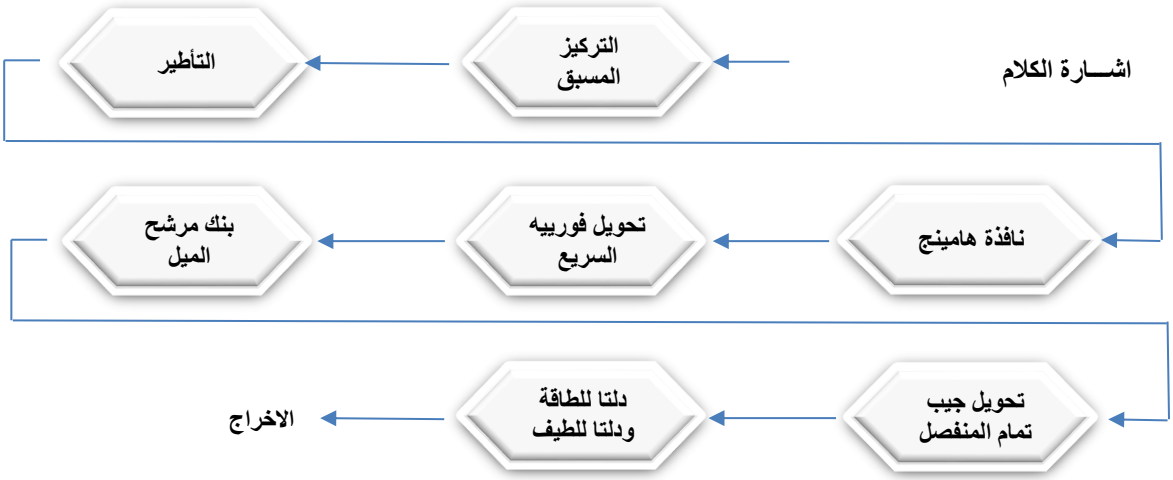
الشكل (1) العلاقة بين مقياس التردد ومقياس الميل [11]

يتم تحديد تردد العتبة لإنتاج مقياس الترددات إلى مقياس ميل هي [11]: للتحويل من

$$f_{mel} = 1127ln\left(1 + \frac{f_{Hz}}{700}\right)\text{----------(1)}$$

حيث F(Mel) هو التردد بالميل و f(Hz) هو التردد العادي بالهرتز، وهذه العلاقة موضحة في الشكل (1).

يعد استخراج واختيار أفضل تمثيل للمعلمات للإشارة الصوتية مهمة في تصميم أي نظام للتعرف على الكلام. توفر مجموعة من معاملات درجة النغم (MFCC) تمثيلًا مضغوطًا، وهو نتيجة تحويل جيب التمام للوغاريتم الحقيقي لطيف الطاقة قصير المدى المعبر عنه على مقياس تردد ميل. أثبتت معاملات درجة النغم (MFCC) أنه أكثر فعالية [11]. يتضمن حساب معاملات درجة النغم (MFCC) ما يلي:



الشكل (2) رسم تخطيطي لخطوات حساب معاملات درجة النغم

كما هو مبين في الشكل (2)، تتكون معاملات درجة النغم من سبع خطوات حسابية. كل خطوة لها وظيفتها وطريقتها الرياضية الخاصة، كما هو موضح أدناه:

3-1-1: التركيز المسبق(Pre-Emphasis)

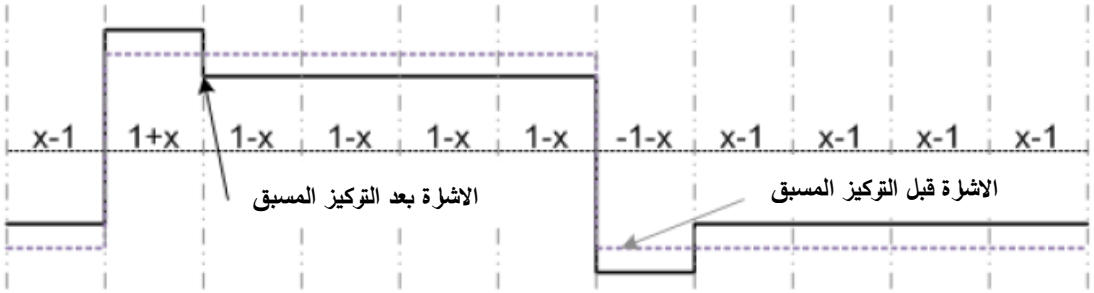
تتناول هذه الخطوة عملية تمرير الإشارة عبر الفلتر (وهو من نوع High Pass Filter)، يؤكد على الترددات الأعلى. ستزيد هذه العملية من طاقة الإشارة عند الترددات الأعلى. وذلك باستخدام الصيغة التالية [13]:

$$Y(n) = X(n) - \alpha * X(n-1) \text{----------(3)}$$

حيث تمثل X(n) تمثل إشارة الادخال.

Y(n)تمثل إشارة الإخراج بعد عملية التركيز المسبق.

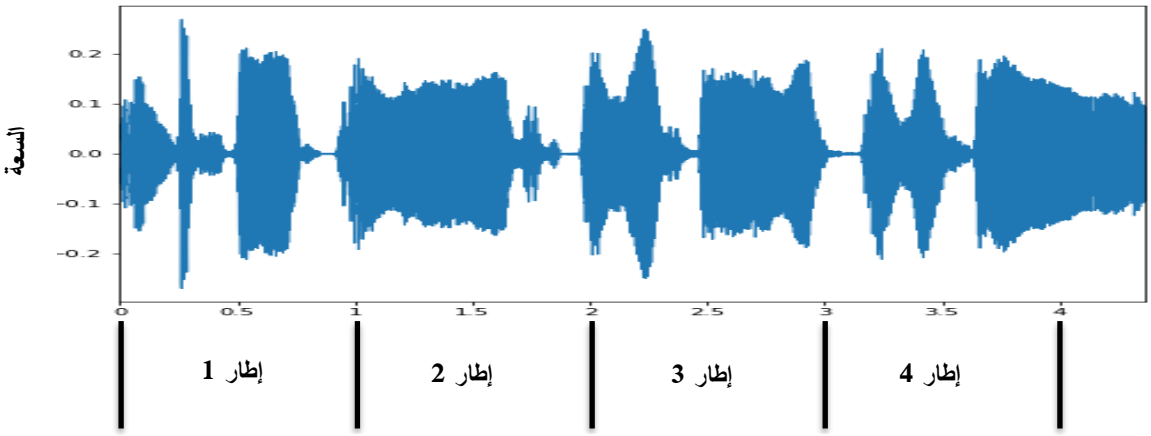
α يمثل ثابت تتراوح قيمته بين (0.9) و(1) .



الشكل (3) يوضح شكل الإشارة قبل وبعد عمل التركيز المسبق

3-1-2: التأطير(Framing)

عملية تجزئة عينات الصوت التي تم الحصول عليها إلى إطارات صغيرة بطول يتراوح من 20 إلى 40 مللي ثانية. تنقسم إشارة الكلام إلى إطارات من عينات .N يتم فصل الإطارات المجاورة بواسطة M بحيث (M<N). القيم النموذجية المستخدمة هي M = 100 و N = 256. مع وجود تداخل اختياري يساوي نصف او ثلث حجم الإطار وذلك من اجل تسهيل الانتقال من إطار الى اخر [11]، كما هو موضح في الشكل (3) ادناه.



الشكل (4) عملية التأطير على الاشارة

3-1-3: نافذة هامينج (Hamming Window)

من خلال النظر في الكتلة التالية في سلسلة معالجة استخراج الميزة ودمج جميع خطوط التردد الأقرب، يتم استخدام نافذة Hamming كشكل للنافذة. معادلة نافذة هامينج هي [13]:

$$W(n) = W_0\left(n - \frac{N-1}{2}\right)\text{----------(4)}$$

إذا عرفت النافذة على أنها $0 \le n \le N-1$ W(n), إذ أن:

N = عدد العينات في كل اطار.

تهدف هذه الخطوة الى إنشاء النافذة في كل إطار فردي لتقليل انقطاع الإشارة في بداية ونهاية كل إطار. إذا حددنا النافذة على أنها $0 \le n \le N-1$ W(n) حيث N هو عدد العينات في كل إطار. لذلك، يمكن عرض نتيجة إنشاء النافذة بناءً على المعادلة التالية [13]:

$$Y(n) = X(n).W(n)\text{----------(5)}$$

Y(n) = إشارة الاخراج.

X(n) = إشارة الادخال.

W(n) = نافذة هامينج.

هنا، يتم استخدام نافذة Hamming بشكل أكثر شيوعًا كشكل النافذة في تقنية التعرف على الكلام، ودمجت جميع خطوط التردد الأقرب من خلال النظر في الكتلة التالية في سلسلة معالجة استخراج الميزة. تظهر الاستجابة النبضية لنافذة هامينج وفقًا للمعادلة التالية [13]:

$$W(n) = 0,54 - 0,46.\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \le n \le N-1\text{----------(6)}$$

لهذا السبب، يتم استخدام نافذة Hamming لاستخراج معاملات درجة النغم، مما يقلل من قيمة الإشارة نحو الصفر عند حدود النافذة ويتجنب الانقطاعات.

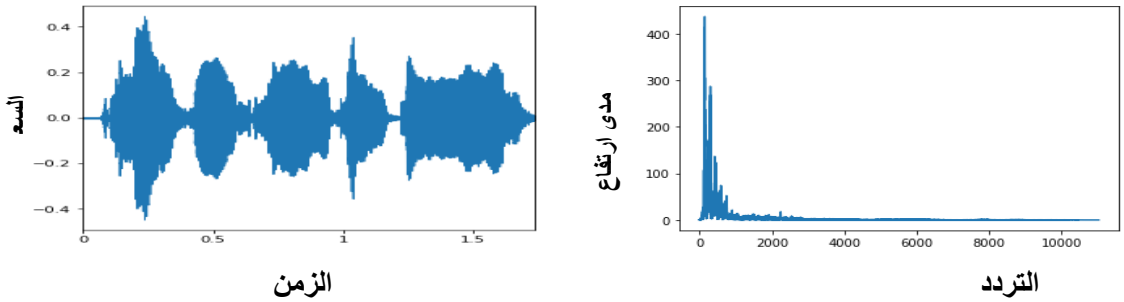
3-1-4: تحويل فوريه السريع (Fast Fourier Transform)

تحويل العينات N لكل إطار في المجال الزمني إلى مجال التردد. تحويل فوريه هو التفاف النبضة المزمنة U[n] والاستجابة النبضية للقناة H[n] في المجال الزمني.

البيان يدعم المعادلة التالية [13]:

$$Y(w) = FFT[h(t) * x(t)] = H(w).\ X(w) \quad \text{----------(7)}$$

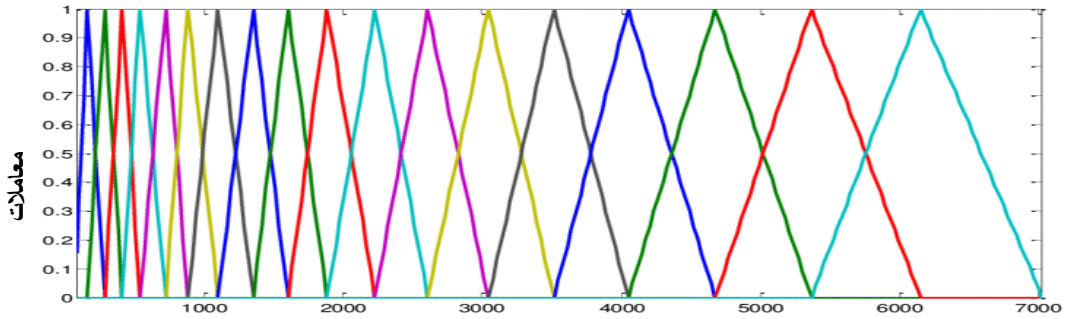
حيث ان Y(w) ، H(w)، X(w) هي تحويل فوريه السريع لكل من h(t) ، x(t) .



الشكل (5) تحول الإشارة من المجال الزمني الى المجال الترددي بعد تطبيق تحويل فوريه السريع

3-1-5: معالجة بنك مرشح ميل (Mel Filter Bank Processing)

نطاق التردد في طيف تحويل فوريه السريع واسع جدًا والإشارة الصوتية لا تتبع مقياسًا خطيًا. فيتم تشغيل بنك المرشح وفقًا لمقياس ميل كما هو موضح في الشكل (5).



الشكل (6) بنك مرشح مقياس ميل [13].

يوضح الشكل مجموعة من المرشحات المثلثية المستخدمة لحساب المجموع المرجح للمكونات الطيفية للمرشح بحيث يكون ناتج العملية مقاربًا لمقياس ميل. تكون استجابة تردد الاتساع لكل مرشح مثلثة، تساوي 1 عند التردد المركزي، وتنخفض خطيًا إلى الصفر عند التردد المركزي لمرشحين متجاورين. لذا فإن ناتج كل مرشح هو مجموع مكوناته الطيفية التي تمت تصفيتها. بعد ذلك، تُستخدم المعادلة التالية لحساب ميل لتردد معين (f) [11]:

$$F(Mel) = 2595.log10\left(1 + \frac{f}{700}\right) \text{----------(8)}$$

3-1-6: تحويل جيب التمام المنفصل (Discrete Cosine Transform)

يوفر تمثيل طيف الكلام تمثيلًا جيدًا للخصائص الطيفية المحلية للإشارة لتحليل إطار معين، نقوم بتحويل طيف طاقات درجة النغم الى المجال الزمني باستخدام تحويل فوريه المنفصل (Discrete Fourier Transform). وتسمى النتيجة بمعاملات درجة النغم

(MFCC)كما موضح بالشكل (6)، تسمى مجموعة المعاملات متجه الصوت (acoustic vector). لذلك، يصبح كل بيان إدخال تسلسل ناقلاً صوتياً [14]:

$$Cn = \sum_{k=1}^{k} (logSk)\cos\left(n.\left(k-\frac{1}{2}\right).\frac{\pi}{k}\right)\text{----------(9)}$$

حيث انn=1,2,……,k

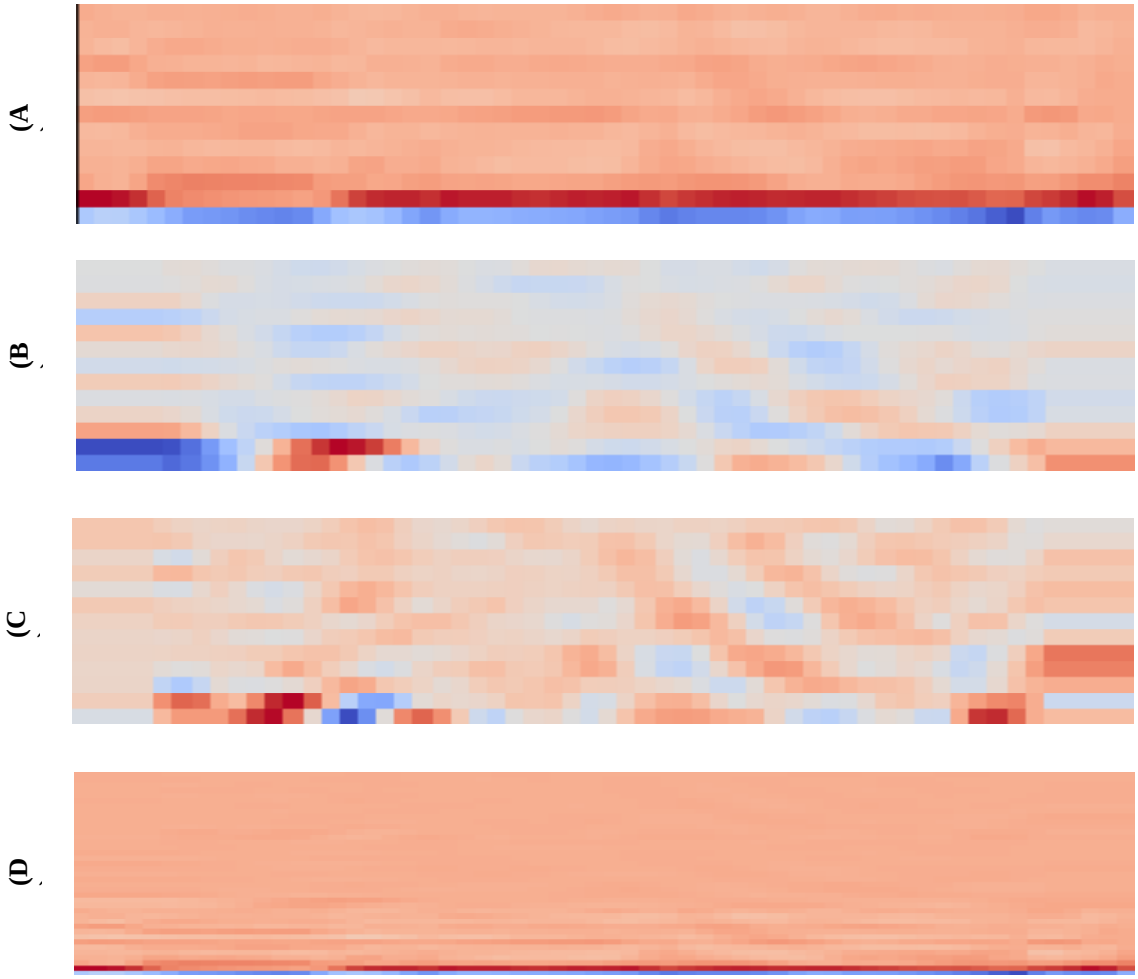
بينما Sk,k=1,2,……,k هي مخرجات آخر خطوة.

3-1-7: دلتا للطاقة ودلتا للطيف (Delta Energy and Delta Spectrum)

ترتبط الطاقة بهوية الصوت، وهي إشارة للكشف عن الصوت. يتم التعبير عن الطاقة في إطار الإشارة x في النافذة من العينة الزمنية t1 إلى العينة الزمنية t2 بالمعادلة التالية [15]:

$$\text{الطاقة}=\sum_{t=t1}^{t2} x^2(t) \quad \text{----------(10)}$$

يمكن أن يتأثر أداء معاملات درجة النغم بتردد ميل بمكونين، الأول هو عدد المرشحات والثاني هو نوع النافذة [14,15،16].



الشكل (7) (A) معاملات درجة النغم. (B) المشتقة الأولى لمعاملات درجة النغم. (C) المشتقة الثانية لمعاملات درجة النغم. (D) دمج معاملات درجة النغم مع مشتقتها الأولى والثانية

4- الشبكات العصبية الالتفافية (Convolutional Neuron Network)

نقدم إطار عمل تصنيف اللغة (Language Classification) استنادًا إلى الشبكات العصبية الالتفافية (CNN). مخطط تعلم الخصائص التمييزية الشبكة العصبية الالتفافية والذي يستخدم مخططات طيفية للإشارة إلى حالة نزاع المتحدث. تحتوي بنية لشبكة العصبية الالتفافية المقترحة ثلاث طبقات التفافية وطبقة متصلة بالكامل يتبعها مصنف (SoftMax). المخطط الطيفي لإشارة الكلام هو تمثيل ثنائي الأبعاد للتردد مقابل الوقت، ويحتوي على معلومات أكثر من النصوص النصية، ويستخدم لتحديد لغة المتحدث. يحتوي المخطط الطيفي على كمية كبيرة من المعلومات التي لا يمكن استخلاصها وتطبيقها عند تحويل إشارات الصوت والكلام إلى نص أو صوتيات. بفضل هذه القدرة، يحسن المخطط الطيفي التعرف على لغة المتكلم. الفكرة الرئيسة هي تعلم الخصائص التمييزية المتقدمة للإشارة الصوتية. هذا هو السبب في أننا نستخدم بنية لشبكة العصبية الالتفافية لتعلم الميزات المتقدمة. المخطط الطيفي مناسب تمامًا لهذه المهمة، استخدمت خصائص معاملات درجة النغم معًا، ويتم استخدام لشبكة العصبية الالتفافية من أجل تصنيف اللغة [16]. كما واستخدمت وظيفة المخطط الطيفي لتحقيق أداء جيد في تصنيف اللغة [17].

5- دوال التنشيط (Activation Functions)

هناك العديد من دوال التنشيط المختلفة. إحدى هذه الدوال الشائعة الاستخدام هي الوحدة الخطية المصححة (Rectified Linear Unit (ReLU)). لقد ثبت أن دالة ReLU تؤدي بشكل عام أداءً أفضل من الخيارات الأخرى وتستخدم على نطاق واسع اليوم 8. يظهر تعريف وظيفة تنشيط relu أدناه في المعادلة 10 [18]:

$$f(x) = \begin{cases} 0 & for\ x < 0 \\ 1 & for\ x \geq 0 \end{cases} \text{----------(11)}$$

دالة التنشيط الأخرى هي (SoftMax)، والتي تُستخدم بشكل شائع لطبقات الإخراج للشبكات العصبية. تأخذ وظيفة (SoftMax) تنشيط جميع الخلايا العصبية n للطبقة وتنشئ توزيعًا احتماليًا يتكون من عدد n من الاحتمالات. مع متجه الإدخال x الذي يحتوي على تنشيط كل خلية عصبية، ستكون دالة (SoftMax)، المشار إليها، متجهًا بنفس طول x تحتوي على الاحتمالات المحسوبة [19]. يتم تعريف التنشيط في الخلايا العصبية i باستخدام وظيفة SoftMax:

$$\sigma(x)_i = \frac{e^{xi}}{\sum_{j=1}^{n} e^{xj}} \text{ ----------(12)}$$

$$i = 1, 2, \dots, n \ and \ x = (x_1, x_2, \dots, x_n) \in R^n$$

6- النظام المقترح (Proposed System)

صمم النظام على مراحل متتابعة تضم المعالجة الأولية على الصوت واستخلاص الصفات باستخدام خوارزمية معاملات درجة النغم (MFCC)، الصفات المستخلصة تم خزنها شكل صور طيفية لخصائص الصوت بالامتداد (.png) على ثم مرحلة بناء النظام المقترح كمصنف (Classifier) باستخدام خوارزمية الشبكة العصبية الالتفافية(CNN) ، ثم تبدأ مرحلة التدريب لحين الوصول الى نتائج جيدة، تخزن المعاملات لكي تستخدم بعملية التنبؤ بلغة المتكلم في مرحلة الاختبار. استخدمت لغة بايثون (Python) في بناء النظام، حيث تعد لغة برمجية سهلة التعلم ومفتوحة المصدر. النظام المقترح يتضمن أربع مراحل أساسية:

- المرحلة الأولى: اعداد قاعدة البيانات للملفات الصوتية للمتكلمين.

- المرحلة الثانية: المعالجة الأولية على الملفات الصوتية.
- المرحلة الثالثة: استخلاص الصفات باستخدام خوارزمية معاملات درجة النغم (MFCC).
- المرحلة الرابعة: بناء المصنف والتنبؤ بلغة المتحدث باستخدام خوارزميات الذاكرة طويلة المدى.

هيكلية النظام المقترح موضحة في الشكل (8)

الشكل (8) هيكلية النظام المقترح

إعداد مجموعة البيانات (Preparing Dataset)

تُعد قواعد البيانات الصوتية أساس أي نظام تصنيف اللغة وتشكل البنية التحتية لنظام التخاطب مع الحاسوب. التحدي الأول لهذا البحث هو كيفية العثور على مجموعة بيانات من مقاطع الصوت بلغات مختلفة كبيرة بما يكفي لتدريب شبكة. واعتمد على قاعدة البيانات (M2L-Dataset) والتي ثلاث لغات وهي (العربية والإنكليزية والكردية) وبمعدل ألف عينة لكل لغة من اللغات الثلاثة من نوع (.WAV) وبمعدل عينة (Sample Rate) مقداره (22050)، تتكون عينات اللغة العربية من تسجيلات صوتية ل 40 شخص من كلا الجنسين بمعدل 25 عينة لكل شخص. اما بالنسبة للغة الكردية فقد تم الحصول عليها من بعض المحاضرات والدروس إذ تم تقطيعها معالجتها بما يناسب قاعدة البيانات من اجل الحصول على عدد عينات كافية للتدريب والاختبار. اما بالنسبة للغة الإنكليزية فقد تم الحصول على العينات الخاصة بهذه اللغة من موقع ((VoxForge. url: voxforge.org) VoxForge) وهو مصدر مفتوح يتكون من مقاطع صوتية للمستخدمين بلغات مختلفة.

جدول (1) تفاصيل قاعدة البيانات

التسلسل	اللغة	عدد العينات
1	العربية	1000 ملف صوتي
2	الإنكليزية	1000 ملف صوتي
3	الكردية	1000 ملف صوتي

تم جمع الصوتية العربية والكردية. كان المقاطع باللغات والإنجليزية المتحدثون أصحاب لهجات مختلفة وكانوا من جنسين مختلفين. قد يتحدث نفس المتحدثين في أكثر من مقطع واحد.

6-1المعالجة الاولية (Preprocessing)

تعد المعالجة المسبقة للبيانات خطوة أساسية من اجل التعرف على لغة المتحدث، لأنها تضمن إعداد البيانات جيدًا لأنواع معينة من التحليل. في هذا البحث، المعالجة المسبقة تمت على خطوتين:

- ازالة فترات الصمت.
- توحيد طول الملفات ب ثانية واحدة او ثانيتين.

6-2استخلاص الصفات (Feature Extraction)

في هذا البحث تم استخدام معاملات درجة النغم (MFCC) من اجل استخلاص الصفات والتي تخزن على شكل صور طيفية باستخدام مكتبة (librosa). القيم المستخدمة في عملية استخلاص الصفات موضحة في الجدول التالي:

الجدول (2) القيم المستخدمة في استخلاص معاملات درجة النغم

اسم الحقل	قيمة الحقل
معدل أخذ العينات (Sampling Rate)	22050
طول القفزة (Hop-length)	512
عدد معاملات درجة النغم	13

6-3التصنيف والاختبار (Classification and Testing)

لتنفيذ عملية التصنيف واكتشاف اللغة تم بناء نموذج لشبكة عصبية التفافية (CNN)، حيث تم تنظيم بيانات الإدخال كسلسلة من خرائط الميزات قبل إدخالها في الشبكة العصبية الالتفافية لتحديد النمط. تنظيم الإدخال كمصفوفة ثنائية الأبعاد (2-D)، حيث يتم تخزين قيم البيكسل في مؤشرات الإحداثيات. يمكن رؤية قيم RGB (الأحمر والأخضر والأزرق) على أنها ثلاث خرائط ميزة ثنائية الأبعاد منفصلة في الصور الفوتوغرافية الملونة. تقوم الشبكات العصبية الالتفافية بتشغيل نافذة صغيرة فوق صورة الإدخال في كل من وقت التدريب والاختبار، بحيث يمكن لأوزان الشبكة خلال هذه النافذة التعلم من مجموعة متنوعة من عناصر بيانات الإدخال، يُشار إلى قرار استخدام الأوزان المتطابقة في كل نقطة من النافذة باسم تقاسم الوزن الكامل. هيكلية النموذج المقترح موضحة في الجدول التالي:

الجدول (3) هيكلية نموذج الشبكة العصبية الالتفافية المستخدم في البحث		
اسم الطبقة	عدد الخلايا في كل طبقة	دالة التنشيط المستخدمة
Hidden Con2D layer	6	Relu
Hidden Con2D layer	16	Relu
Hidden Con2D layer	64	Relu
Hidden Dense layer	128	Relu
Hidden Dense layer	2 او 3 حسب عدد اللغات	Softmax
عدد دورات التدريب = 50		

فيما يلي تفاصيل تقسيم قاعدة البيانات بعد معالجتها موضحة بالجدول التالي:

جدول (4) استخدام الملفات الصوتية		
مجموع الملفات الكلي المستخدم في البحث هو 3000		
80% من المجموع الكلي قُسم كالاتي		20% من المجمع الكلي لعملية الاختبار
20% للتقييم	80% للتدريب	

7- النتائج والمناقشة

المقياس الذي اعتمد عليه في هذا البحث هو الدقة (Accuracy) اما التأكد من صحة النظام فقد تم الاعتماد على معايير التقييم وهي الدقة (Precision) والاستدعاء (Recall) ودرجة F1 (F1 Score) وهذه القيم يمكن الحصول عليها من مصفوفة الارتباك كما هي موضحة في الجدول التالي:

		الفئة المتوقعة	
		موجب	سالب
الفئة الفعلية	موجب	إيجابي صحيح	سلبي صحيح
	سالب	إيجابي خاطئ	سلبي خاطئ

دقة التصنيف (Accuracy) = عدد التوقعات الصحيحة من بيانات الاختبار / عدد بيانات الاختبار الكلي **----------(13)**

الدقة (Precision) = الايجابيات الصحيحة / (الإيجابيات الصحيحة + الإيجابيات الخاطئة)-------(14)

الاستدعاء (Recall)= الايجابيات الصحيحة / (الإيجابيات الصحيحة + السلبيات الخاطئة)----------(15)

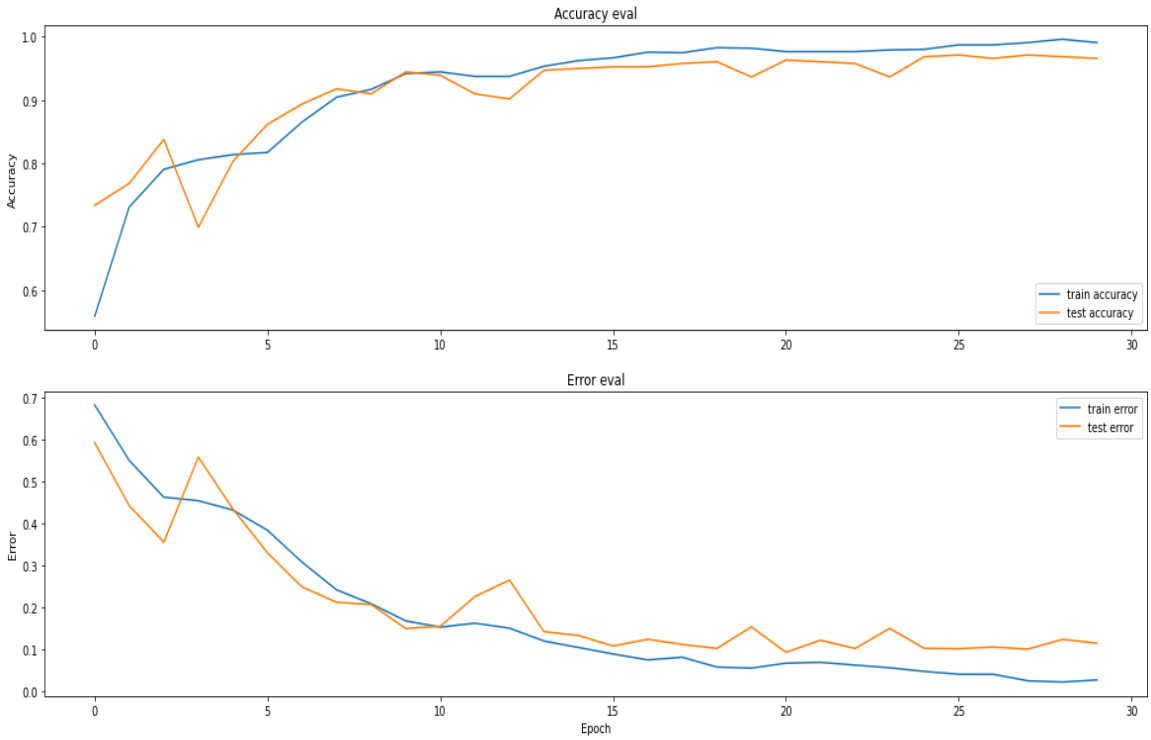
درجة F1 (F1Score) =2* ((الدقة * الاستدعاء) / (الدقة + الاستدعاء))----------(16)

الجدول التالي يوضح النتائج النهائية لدقة التصنيف.

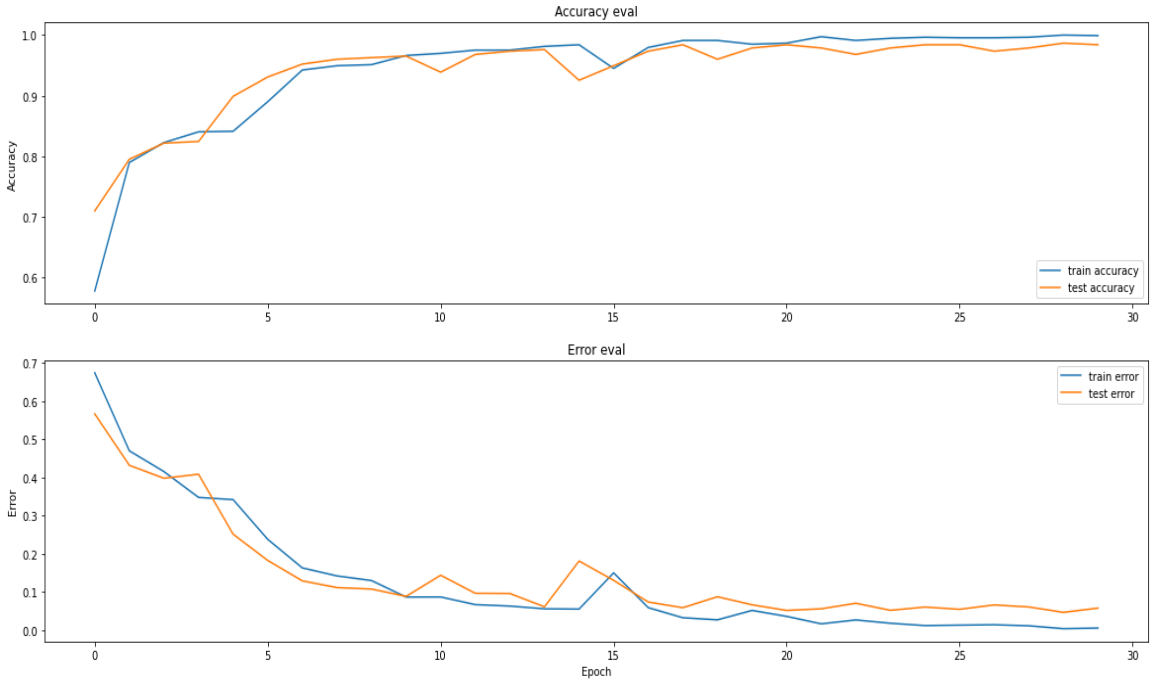
جدول (5) النتائج النهائية لدقة التصنيف.			
التسلسل	عدد اللغات	طول العينة	دقة التصنيف
1	2	1 ثانية	%97.60
2	2	2 ثانية	%98.40
3	3	1 ثانية	%94.26
4	3	2 ثانية	%94.66

اما الجدول التالي قيم معايير التقييم المستخدمة في النظام للحالات الثلاثة للتصنيف.

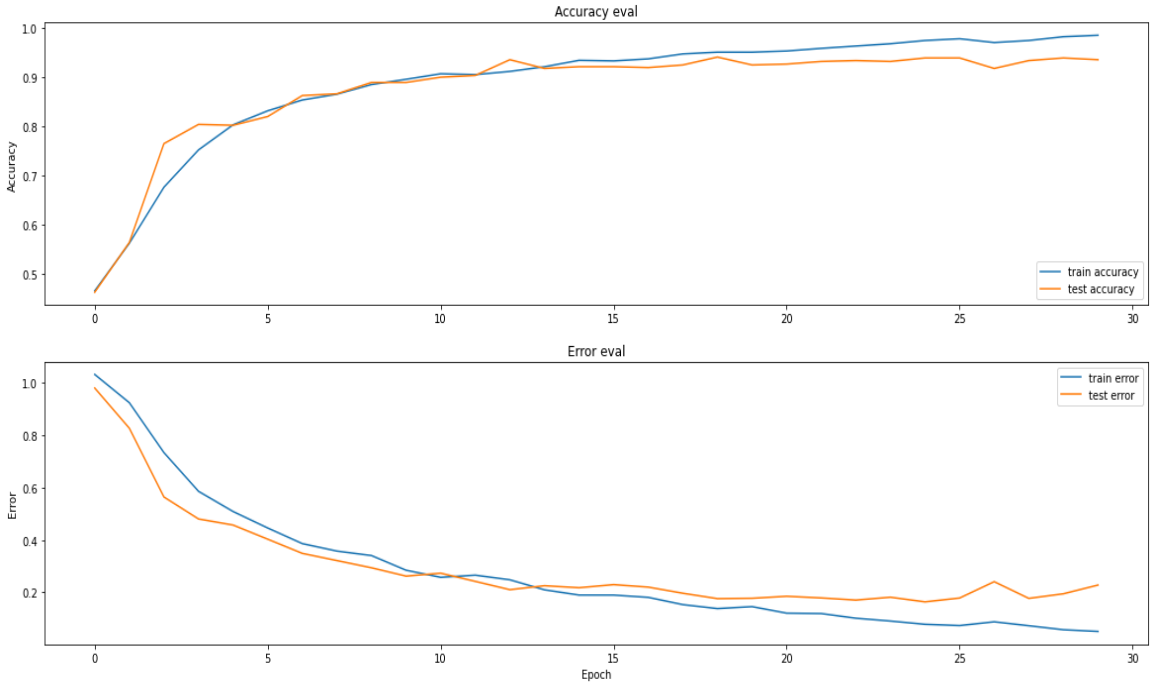
الجدول (6) قيم المعايير المستخدمة في النظام					
(A) قيم المعايير إذا كان طول العينة ثانية واحد والتمييز بين لغتين					
التسلسل	اللغة	support	f1-score	recall	Precision
1	العربية	244	0.98	0.99	0.96
2	الانكليزية	256	0.98	0.96	0.99
(B) قيم المعايير إذا كان طول العينة ثانيتين والتمييز بين لغتين					
التسلسل	اللغة	support	f1-score	recall	Precision
1	العربية	244	0.98	0.98	0.99
2	الانكليزية	256	0.98	0.99	0.98
(C) قيم المعايير إذا كان طول العينة ثانية واحد والتمييز بين ثلاث لغات					
التسلسل	اللغة	support	f1-score	recall	Precision
1	العربية	250	0.94	0.96	0.92
2	الانكليزية	240	0.95	0.97	0.93
3	الكردية	260	0.94	0.90	0.98
(D) قيم المعايير إذا كان طول العينة ثانيتين والتمييز بين ثلاث لغات					
التسلسل	اللغة	support	f1-score	recall	Precision
1	العربية	250	0.94	0.92	0.96
2	الانكليزية	240	0.96	0.97	0.94
3	الكردية	260	0.95	0.95	0.95



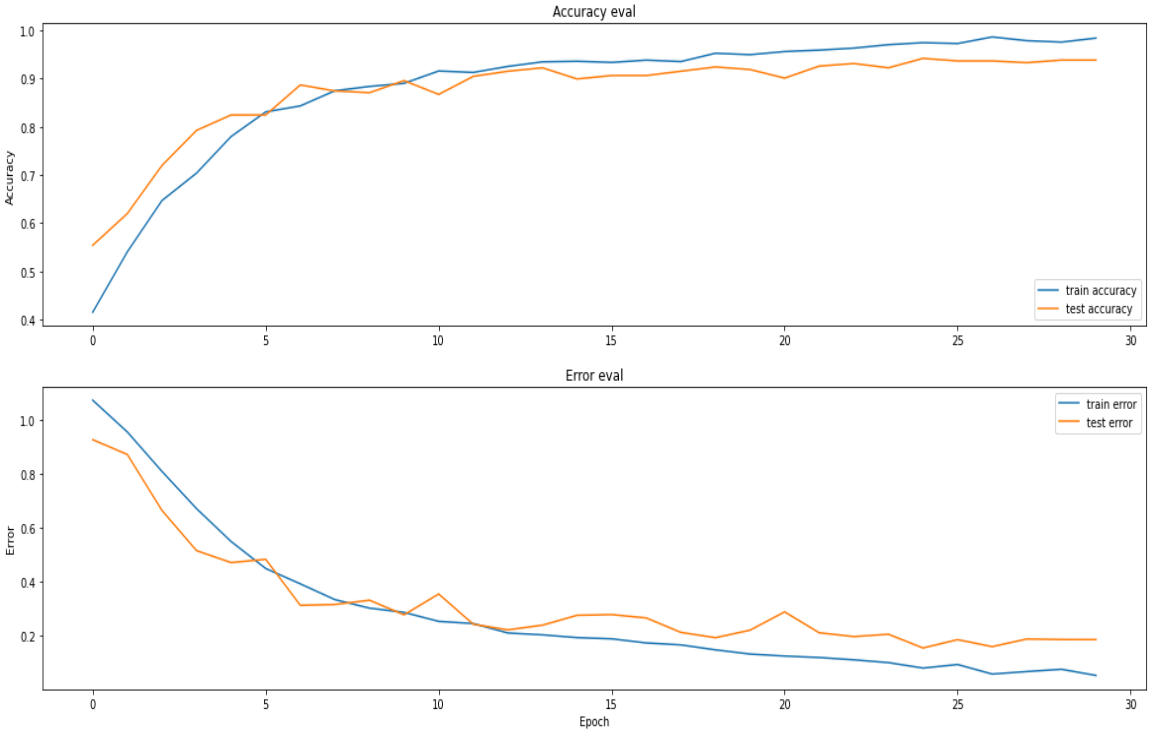
الشكل (9) مخطط تقييم الدقة والخطأ إذا كان طول العينة ثانية واحد والتمييز بين لغتين



الشكل (10) مخطط تقييم الدقة والخطأ إذا كان طول العينة ثانيتين والتمييز بين لغتين



الشكل (11) مخطط تقييم الدقة والخطأ إذا كان طول العينة ثانية واحدة والتمييز بين ثلاث لغات



الشكل (12) مخطط تقييم الدقة والخطأ إذا كان طول العينة ثانيتين والتمييز بين ثلاث لغات

الجدول (7) مقارنة النموذج المقترح مع نماذج الاعمال ذات الصلة بالنظام

اسم الباحث	قاعدة البيانات	استخراج الصفات	الخوارزمية المستخدمة	طول العينة	عدد اللغات	الدقة
Yaakov et al [6]	The OGI Multi-language Telephone Speech Corpus	RASTA Cepstrum Spectrum	J48 Random Forest MultiBoostab BayesNet Logistic Regression Sequential Minimal Optimization	لم تذكر	2	89.18%
					5	81.85%
					7	80.33%
Shauna et al [5]	VoxForge	مخططات طيفية	CNN	4 ثواني	2	%97
					6	%89
Alexandra [7]	مستودع الاخبار الاوربية	مخططات طيفية	CNN CRNN	لم تذكر	4	%82
					6	%71
Anna [8]	تغريدات تويتر	Bag of Words (BoW) N-grams	LSTM	ام تذكر	9	95%
Herman [9]	L5K Music 6 Corpus	MFCC	VGGish	لم تذكر	6	%35.2
النظام المقترح	M2L-Dataset	MFCC	CNN	ثانية	2	%97.60
					3	%94.26
				ثانيتين	2	%98.40
					3	%94.66

8- الاستنتاج

تحقق الهدف المتمثل في إنشاء نموذج قادر على التمييز بين لغتين بدقة 97٪ إذا كان طول العينة ثانية واحدة بينما بلغت دقة التمييز بين لغتين إذا كان طول العينة ثانيتين 98%، كما وكانت دقة النموذج النهائي للتمييز بين ثلاث لغات 95٪ اذا كان طول العينة ثانية واحدة في حين بلغت دقة التمييز بين لغتين اذا كان طول العينة ثانيتين 96%، باستخدام طبقتين كثيفتين وثلاث طبقات التفافية. كان أفضل إعداد فيما يتعلق بمعالجة الإشارات هو استخدام معاملات درجة النغم (MFCC) واستخدام 13 مرشحاً من بنوك الترشيح إذا تم تنفيذه في تطبيق التحكم الصوتي، اقترحت عينة بطول ثانية واحدة او ثانيتين. تم من خلال العمل المقترح استنتاج أيضًا إلى أنه كان من الممكن الوصول إلى نتيجة أفضل إذا كان لدينا المزيد من القوة الحسابية نظرًا لأن الاختبارات التي تستغرق وقتًا طويلاً أدت إلى إجراء عدد أقل من الاختبارات. كان استخدام قاعدة بيانات (M2L-Dataset) مفيداً جدا وذلك لان المواد الصوتية التي استخدمت كانت ذات جودة جيدة، أخيرًا، سيكون من المثير للاهتمام بالنسبة للمشروعات المستقبلية النظر في تنفيذ المزيد من اللغات.

المصادر

[1] Grimes, B, Ethnologue: Languages of the World 18th ed., Dallas: SIL International 2015.
[2] Adarsh.D.Patil, Akshay Vishwas Joshi, Harsha.K.C, Pramod.N, Spoken Language Identification Using Machine Learning, Department Of Computer Science & Engineering M.S.Ramaiah Institute Of Technology(Autonomous Institute, Affiliated To Vtu) Bangalore-560054,Www.Msrit.Edu,May 2012
[3] Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3856–3866.
[4] Bae, J.; Kim, D.-S. End-to-End Speech Command Recognition with Capsule Network. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 776–780.
[5] Revay, Shauna, and Matthew Teschke. "Multiclass language identification using deep learning on spectral images of audio signals." arXiv preprint arXiv:1905.04348 (2019).
[6] HaCohen-Kerner, Yaakov, and Ruben Hagege. "Automatic classification of spoken languages using diverse acoustic features." Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters. 2015.
[7] Draghici, Alexandra, Jakob Abeßer, and Hanna Lukashevich. "A study on spoken language identification using deep neural networks." Proceedings of the 15th International Conference on Audio Mostly. 2020.
[8] Avenberg, Anna. "Automatic language identification of short texts." (2020).
[9] Groenbroek, Herman. A Machine Learning Approach to Automatic Language Identification of Vocals in Music. Diss. 2021.
[10] Cabañas-Molero, Pablo-Antonio. "Classification and separation techniques based on fundamental frequency for speech enhancement." (2016).
[11] Abhishek Manoj Sharma, Speaker Recognition Using Machine Learning Techniques, SJSU ScholarWorks,20/5/2019.

[12] Lindgren, Andreas, and Gustav Lind. "Language Classification Using Neural Networks." (2019).
[13] Bezoui, Mouaz, Abdelmajid Elmoutaouakkil, and Abderrahim Beni-hssane. "Feature extraction of some Quranic recitation using mel-frequency cepstral coeficients (MFCC)." 2016 5th international conference on multimedia computing and systems (ICMCS). IEEE, 2016.
[14] Chapaneri, Santosh. "Spoken digits Recognition using weighted MFCC and improved Features for dynamic time wrapping.", published in International Journal of Computer Applications (0975 – 8887) Volume 40– No.3, February 2012.
[15] S. Singh, and E. Rajan, "Vector Quantization approach for speaker recognition using MFCC and inverted MFCC", International Journal of Computer Applications, vol. 17, no. 1, Mar 2011.
[16] Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Kwon, S.; Baik, S.W. Deep features-based speech emotion recognition for smart affective services. Multimed. Tools Appl. 2019, 78, 5571–5589.
[17] Yu, D.; Seltzer, M.L.; Li, J.; Huang, J.-T.; Seide, F. Feature learning in deep neural networks-studies on speech recognition tasks. arXiv 2013, arXiv:1301.3605.
[18] Dahl, G, Sainath, T, and Hinton, G, improving deep neural networks for LVCSR using rectified linear units and dropout, I International Conference on Acoustics, Speech and Signal Processing, Vancouver, 26 May 2013
[19] Medium Corporation, Understand the SoftMax Function in Minutes, Medium 2018, accessed 18 May 2019,