Journal of Education and Science (ISSN 1812-125X), Vol: 30, No: 4, 2021 (141-160)

Text dependent speaker identification system based on deep learning

Yusra Faisal Al-Irahyim[1*], Qasim Sadiq Mahmood [2*]

[1*,2] Department of Computer Science, Collage of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

E-mail: [1*]yusrafaisalcs@uomosul.edu.iq, [2]Qasim.csp56@student.uomosul.edu.iq

(Received May 16, 2021; Accepted July 03, 2021; Available online September 01, 2021)

DOI: 10.33899/edusj.2021.130144.1161, © 2021, College of Education for Pure Science, University of Mosul.

ABSTRACT

Speaker identification techniques are one of those most advanced modern technologies and there are many different systems had been developed, from methods that used to extract characteristics and classification. The applications of Speech identification are quite difficult and requires modern technologies with a large number of audio samples and resources.

In this research, the system of speaker identification had been designed based on a text (the word or sentences are pre-defined) which give the system the capability to identify the speaker in the least time, number of training samples and resources. The system consists four main parts, the first one is to create audio databases. In the study, two audio databases were relied upon, the first being a database (QS-Dataset) and the second database (audioMNIST_meta). The databases were processed and configured in a way that was explained in the body of the research later. The second part of the research is to extract the characteristics through the pitch coefficients algorithm, while the third part is the use of the neural network as a classifier. And the last part of the research is to verify the work and results of the system.

The test results showed the ability of the MNN network to deal with the smallest number of data, as it achieved a percentage of 100%. As for large data, it ranged from 80% to 81%. Unlike CNN network, the results were not good for the few data, from 60% to 76%, and with large data it was The results are excellent, from 91% to 96%.

Keywords: Speaker identification, MFCC, Multilayer Neural Network (MNN), convolution neural network (CNN), Deep Learning

نظام تحديد المتحدث المعتمد على النص بالاعتماد على التعلم العميق

يسرى فيصل الرحيم ، [2*] قاسم صديق محمود [1*]

قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

الخلاصة

تعد تقنيات التعرف على المتحدث من أهم التقنيات الحديثة وقد طورت العديد من الأنظمة المختلفة من حيث الطرق المستخدمة في استخراج الخصائص وطرق التصنيف. تشكل تطبيقات التعرف على المتحدث تحديًا كبيرًا، حيث يتطلب تقنيات حديثة وعدد كبير من عينات الصوتية وموارد كبيرة.

في هذه البحث صمم نظام التعرف على المتحدث المعتمد على النص (تكون الكلمة او الجمل محددة مسبقا)حيث يتميز النظام بقدرته على التعرف على المتحدث باقل وقت واقل عدد من عينات التدريب واقل الموارد. يتكون النظام من اربعة اجزاء رئيسة وهي: الجزء الأول بناء قواعد بيانات صوتية, واعتمد في الدراسة على قاعدتين للبيانات الصوتية, الاولى قاعدة بيانات (QS-Dataset) وقاعدة البيانات الثانية(audioMNIST_meta).،وقد تم معالجة و تهيئة قواعد البيانات بطريقة تم شرحها في متن البحث لاحقا. والجزء الثاني من البحث استخلاص الخصائص عن طريق خوارزمية معاملات درجة النغم اما الجزء الثالث فهو استخدام الشبكة العصبية كمصنف. واما الجزء الاخير في البحث فهو التحقق من عمل و نتائج النظام.

اوضحت نتائج الاختبار قدرة شبكة MNN على التعامل مع اقل عدد من البيانات حيث حققت نسبة (100%) واما مع البيانات الكبيرة تتراوح من 80% الى 81% على عكس شبكةCNN كانت نتائج غير جيدة بالنسبة للبيانات القليلة من60% الى 76 % ومع بيانات الكبيرة كانت النتائج ممتازة من91% الى 96 % .

الكلمات المفتاحية: التعرف على هوية المتحدث ،MFCC ، الشبكة العصبية متعددة الطبقات (MNN)، الشبكة العصبية الالتفافية (CNN)، التعلم العميق.

1.المقدمة

الصوت عبارة عن إشارة مصنوعة من نغمة (tone) أو عدد من النغمات المتصلة ببعضها البعض لتعني شيئًا ما وتستخدم للتواصل بين البشر أو أي كائن حي ، حيث يعبرون من خلالها عما يريدون قولاً أو فعلاً بوعي أو بغير وعي ، والإحساس الناجم عن ذلك موجات تسمى السمع. بسبب الصوت ، يحصل الناس على العديد من الخبرات في الحياة. في الماضي ، لم يكن الصوت الذي يصدره الإنسان هو الطريقة الوحيدة المستخدمة للتواصل مع بعضهم البعض ، ولكنهم استخدموا أيضًا العديد من الأشياء التي تصدر ضوضاء واهتزازات مثل الطبول والمزامير [1].

وفقًا لـ M. Gray ، تم اقتراح الأفكار الأولى لاستخدام أنظمة التعرف على المتحدث في عام 1966 بواسطة S. Saito و F. Itakura من NTT. ثم في عام 1999 ، طورت إريكسون نظام التعرف على المتحدث لسلسلة هواتفها المحمولة. في نوفمبر 2010 ، استحوذت شركة Nuance Communications على شركة .PerSayلتطوير طريقة جديدة تعتمد على الخصائص البيومترية[2].

يمكن تقسيم تقنية التعرف على المتحدث تقريبًا إلى مجالين فرعيين ، وهما التعرف على الكلام والتعرف على المتحدث [3]. التعرف على الكلام هو نهج لتحليل محتويات الكلمات / الكلام الذي يتحدث به المتحدث. يستخدم كل نظام من أنظمة التعرف على الكلام العديد من الخوارزميات لتحويل الموجات الصوتية إلى بيانات مفيدة للمعالجة والتي يتم تفسيرها بعد ذلك بواسطة الجهاز. هذه الأنظمة تنتج بعد ذلك مخرجات تم إنشاؤها في شكل نص ليتم استخدامه [4]. حيث أن التعرف على المتحدث هو الطريقة المستخدمة لتحديد عندما يتحدث المتحدث بهذه الكلمات / الكلام. تحاول تقنيات التعرف على المتحدث تغطية الجوانب المختلفة للتعرف على الأشخاص من خلال أصواتهم. لأن كل متحدث له طريقته المميزة في التحدث ، بما في ذلك استخدام لهجة معينة ، والإيقاع ، وأسلوب التنغيم ، ونمط النطق

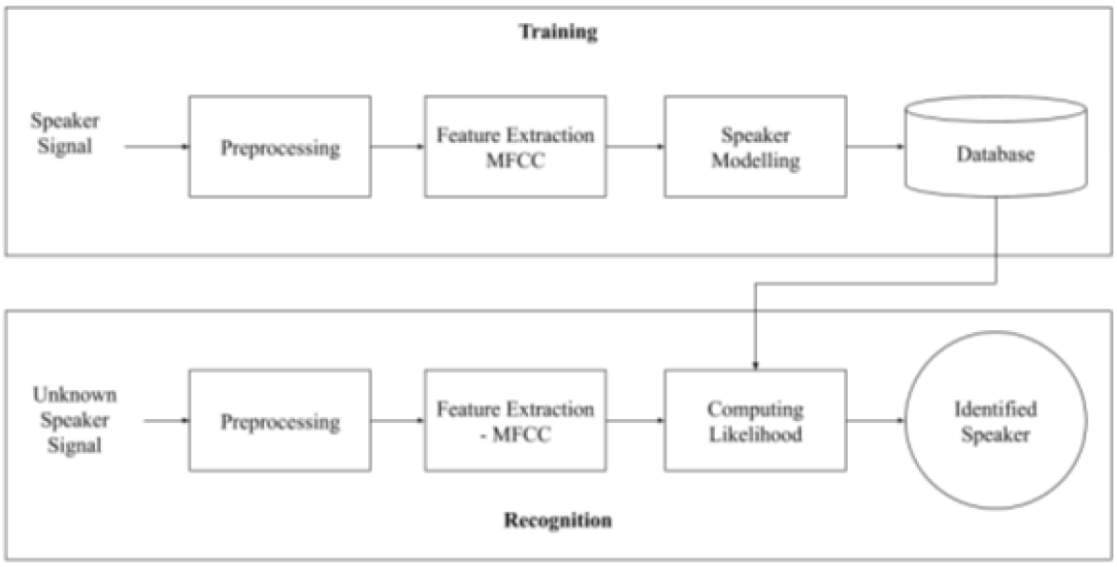
، واختيار المفردات ، وما إلى ذلك، يمكن تقسيم أنظمة التعرف على المتحدث إلى نوعين المعتمدة على النص والمستقلة عن النص. في الأنظمة المعتمدة على تكون عبارات التعرف ثابتة أو معروفة مسبقًا ،وفي الأنظمة المستقلة عن النص ، لا توجد قيود على الكلمات التي يُسمح للمتحدثين التحدث بها[5].

تتمثل الصعوبة الرئيسية لنظام التعرف على المتحدث في أنه من المستحيل نطق كلمة بالطريقة نفسها تمامًا في وقتين مختلفين. يعتمد ذلك على مدى سرعة نطق الكلمة ، ويمكن أن تختلف النغمة [2].

إنّ نظام تمييز المتحدث هو نظام يعتمد على الكلام إذ يتم تحليل إشارة الكلام للحصول على ميزات أقل تنوعًا وأكثر تمييزًا ، ولاستخراج هذه القيم يتم استخدام الطرق مختلفة تسمى ميزات إشارة الكلام (Feature extraction)، ويتم استخدام الميزات التي تم الحصول عليها من إشارة الكلام لإنشاء أنماط المتحدث (Patterns) التي تحتوي على معلومات مختلفة من الكلام، وعادةً ما يتم تحقيق مهمة التعرف على المتحدث من خلال الحصول على إشارة الكلام واستخراج الميزة و نمذجة ميزات الكلام ومطابقة الأنماط والحصول على درجة المطابقة المطلوبة. وتُعد الطرق التعلم العميق (Deep learning) أحد أهم التقنيات المستخدمة لتصميم مختلف أنظمة تمييز الأنماط في مجال معالجة إشارة الصوت.

الهدف من هذه الدراسة هو بناء نموذج نظام يمكن استخدامه في عديد من تطبيقات (البنوك وبالانظمة الأمنية ...الخ) حيث يقوم النظام بالتعرف على المتحدثين من خلال الخصائص الفيزيائية الفريدة لكل صوت وتنتقل من الطرق التقليدية إلى الطرق أكثر صرامة وموثوقية في تحديد هوية المتحدث. لهذا السبب، تلعب المعالجة الرقمية لإشارة الكلام والطريقة التي سيتعامل بها النظام مع الأصوات الواردة وإجراءات تحليل البيانات دورًا مهمًا في التعرف السريع على المتحدث. من أجل التعرف على المتحدث في نظام تحديد المتحدث ، يجب أن تمر الإشارة عبر عدة مراحل منها كما في الشكل (1):

1. تسجيل و معالجة الاشارة
2. استخراج الخصائص
3. مطابقة الخصائص



الشكل(1): الهيكل الأساسي لنظام التعرف على المتحدث [6]

ويتكون البحث من الفقرات الرئيسية التالية:

1. الدراسات السابقة
2. انظمه تحديد هوي المتحدث
3. قاعدة البيانات
4. معالجة الإشارة الرقمية
5. استخراج الخصائص الصوتية
6. ملف القاموس Code_Dictionary
7. الشبكات العصبية الاصطناعية
8. الشبكات العصبية العميقة
9. بناء النظام
10. مقاييس النتائج والتحقق منها
11. النتائج والمناقشة
12. الاستنتاجات

2.الدراسات السابقة

جذبت تقنيات التعرف على المتحدث العديد من الباحثين على مدار الاعوام الماضية وذلك لما لها من تطبيقات مهمة في مختلف المجالات، أن التعرف على المتحدث تعد أحد اهم الخطوات في العديد من التطبيقات منها تحسين التعرف على المتحدث في نظام البنوك وتطبيقات التعرف الالي للمتكلم وغيرها، ولا سيما في وجود التطورات التكنولوجية الحديثة ومعالجة الاشارة والتعلم الالي، وفيما يأتي استعراض لعدد من الدراسات التي قدمها الباحثون في مجال التعرف:

- في سنة 2015 قام كل من Diana Mulitaru و Inge Gavant بدراسة المصنفات المختلفة في التعرف سواء التعرف على الكلام او التعرف على المتحدث.في هذه الدراسة تم الاعتماد على قاعدة بيانات المتاحه تجاريا (TIMIT) لفحص المصنف،كما تم استخلاص الخصائص عن طريق دمج خوارزمتي التنبؤ الخطي الإدراكي (perceptual linear prediction(PLP))و درجة النغم (MFCC) ،حيث أظهرت النتائج هذه الدراسة تفوق الشبكات العصبية على نموذج ماركوف المخفي بنسبة 10% [6].
- في عام 2016 قام كل من Angali Garg و Poonam Sharma بدارسة تبين معدل التعرف على كلمات اللغة الهندية وتأثير الجنس ونوع الشبكة العصبية وخوارزمية استخلاص الخصائص على نسبة التعرف. تبين في الدارسة أن نسبة التعرف على كلام الاناث تفوق مقابلتها للذكور بنسبة 2% كما أن عملية دمج الخصائص PLP مع MFCC أدت الى تحسين نسبة 19% مقارنة مع PLP و 1% مقارنة مع MFCC[7].
- في سنة 2018 استعمل الباحثان Anett Antonyو R.Gopikakumari طريقة جديدة للتعرف على المتحدث المعتمد على النص و الغير معتمد على النص بستخدام الشبكة بيرسبترون متعددة الطبقات ((Multi−layer perceptron (MLP) كاداة تصنيف ،حيث تتميز باقل تعقيد. تم تقليل تعقيد الشبكة من خلال استخراج الخصائص للإشارة الصوتية عن طريث دمج خوارزمية درجة النغم (MFCC) و خوارزمية التحويل الحقيقي الفريد من نوعها (Unique mapped real transform (UMRT)) .توصلت الدراسة انه عند دمج الخوارزميات تكون نتائج استخراج الخصائص افضل بكثير مما إذا استخرجت

الخصائص عن طريق فقط خوارزمية درجة النغم (MFCC).واستخدمة في هذه الدراسة قاعدة بيانات مكونة من 7 انات و8 ذكور.واظهرات النتائج دقة الشبكة المعتمد على النص بنسبة 97.91% و الغير معتمد على النص كانت النسبة 94.44% [8].

- في سنة 2018 قام الباحث Vidya Thanda Setty باستخدام طريقة تهدف للتعرف على المتحدث من خلال شبكة العصبية العميقة ((Deep neural network (DNN).تقوم هذه الدراسة الى تقليل التعقيد الموجود في شبكة العصبية العميقة (DNN) التي قد تؤثر على أداء(من ناحية سرعة التدريب و عدد كبير من عينات الصوتية)، وذلك عن طريق تقليل المعاملات دون التسبب في خسارة كبيرة في الأداء.يتم تقليل المعاملات عن طريق تقليل عدد الطبقات المخفية و عدد الخلاية العصبية الموجودة في كل الطبقات من خلال تصميم دقيق لهذه الشبكة.وتم فحص الشبكة عن طريق قواعد البيانات المتاحة تجاريا (TIMIT وHTIMIT) المكونة من 15 الى 20 شخضا. وظهرت النتائج دقة الشبكة التي استخدمت قاعدة البيانات TIMIT بنسبة 100 % التي تتميز بوضوح الصوت ، اما قاعدة البيانات HTIMIT كانت بنسبة 96.75% ،حيث تتكون من عينات صوتية مسجلة عن طريق الهاتف [9].
- وفي عام 2018 قام كلٌ من Safavi واخرون بالتركيز على التعرف على المتحدث والجنس والفئة العمرية من خلال كلام الأطفال، حيث تم مقارنة أداء العديد من طرق التصنيف مثل نموذج الخليط الغاوسي مع نموذج الخلفية العالمية (GMM–Universal Background Model (GMM–UBM) ، وآلة ناقلات الدعم لنموذج الخليط الغاوسي والمناهج المعتمدة على i-vector. وأظهرت النتائج أنه بالنسبة للتعرف على المتحدث ينخفض معدل الخطأ مع زيادة العمر أما بالنسبة للتعرف على الجنس يكون تأثير العمر أكثر تعقيدًا ويرجع ذلك أساسًا إلى تأثيرات ظهور سن البلوغ. كما أكدت النتائج على أهمية التركيز على خصائص الكلام للأطفال في الاعتبار عند تصميم أنظمة التعرف التلقائي على المتكلم والجنس والفئة العمرية ، وتشمل هذه الخصائص الترددات الأعلى للإشارة وتأثير التردد الأساسي الأعلى على استخراج الميزة وتأثيرات سن البلوغ [10].
- في عام 2019 تناول كلٌ من Jagiasi و اخرون مشكلة نظام التعرف على المتحدث في الجهزة الذكية اذا ما تم استخدام هذه ن في المستقبل.قدمت هذه الدراسة بعض التحسينات التي ممكن ان تحسن الأجهزة الذكية وتكون لديها قدرة عالية للتعرف على المتحدث ،وذلك عن طريق بناء نظام قائم على استخراج الخصائص من إشارة الكلام التي تعود للمتحدث عن طريق خوارزمية معاملات درجة النغم ((Mel Frequency (MFCCs Cepstral Coefficients) المطبقة على الشبكة العصبية الالتفافية (CNN)كأداة تصنيف. وفي هذه الدراسة تم الاعتماد على قاعدة بيانات مكونة من قبل الباحثين،حيث تتضمن عينات صوتية تعود الى 50 شخصا من الذكور والانات و بلغات مختلفة (الإنكليزية والهندية و الماراثي). واظهرت النتائج الدقة للشبكة العصبية الالتفافية 75–80 %،وكما اقترحت هذه الدراسة الى اجراء المزيد من الدراسات وكذلك زيادة عدد الأشخاص في عملية تدريب الشبكة [11].
- في عام 2020 قامت Abd El-Moneim وآخرون بدراسة التعرف على المتحدث في ظل ظروف مختلفة من حيث الضوضاء والصدى، واستخدمت مجموعة فرعية من مجموعة بيانات الماندرين الصينية التي تضمنت تسجيل صوتي لخمس متحدثات ينطقن مئة كلمة مختلفة لكل متحدثة، وتم اختيار 70 كلمة للتدريب و 30 كلمة للاختبار. كما استخدمت طريقتي معاملات درجة النغم(MFCC) والطيف اللوغاريتمي كأدوات لاستخراج صفات الصوت، وعولجت هذه الميزات باستخدام الشبكة العصبية المتكررة للذاكرة طويلة–قصيرة المدى (LSTM–RNN) كأداة تصنيف. أظهرت النتائج أن معدل التعرف على المتحدث للكلام غير المشوه يصل إلى 95.33٪ باستخدام MFCCs، بينما يزداد إلى 98.7٪ عند استخدام الطيف

اللوغاريتمي، بينما أظهرت النتائج بالنسبة للكلام المشوه أن معدل التعرف على المتحدث يصل إلى 90% باستخدام طريقة الطيف [12].

اثبتت الدراسات السابقة ان بناء أي نظام للتعرف على المتحدث يعتمد بشكل أساسي على احدى خوارزميات (MFCC ,PLP,UMRT) او الدمج بين هذه الخوارزميات لاستخلاص الخصائص واحدى الشبكة العصبية (DNN,CNN,RNN) كاداة تصنيف.كما توصل الباحث انه لا يمكن الجزم بوجود خوارزمية افضل من غيرها لان دقة التعرف ترتبط بطبيعة اللغة والجنس المتحدث وعمره والكثير من العوامل الأخرى، كما ان المصنف يلعب دوراً كبيراً في نسبة الدقة التعرف ،لذلك لايمكن اعتماد نتائج البحث ما كنتائج شاملة وذلك بسبب التنوع الكبير جدا في ظروف عملية التعرف ومحدداتها.

بناء على الدراسات السابقة تم اختيار الشبكة العصبية الالتفافية (CNN) و شبكة العصبية متعددة الطبقات (MNN) كمصنف ،و تم اختيار خوارزمية درجة النغم (MFCC) في استخلاص الخصائص.

3.أنظمة تحديد هوية المتحدث

تتطلب أنظمة التعرف على هوية المتحدث عادةً العديد من العمليات والمراحل التي يجب أن تمر فيها الإشارة الصوتية للوصول إلى النتيجة. هناك فئتان من هذه الأنظمة التي تستخدم للتعرف على هوية المتحدث منها الانظمة المستقلة عن النص أو الأنظمة التي تعتمد على النص.

في الأنظمة التي تعتمد على النص، حيث تكون الكلمة او الجمل معروفة مسبقا. في الأنظمة المستقلة عن النص، لا توجد قيود على الكلمات او الجمل التي يُسمح للمتحدثين التحدث بها. يعتبر التعرف المستقل على النص هو التحدي الأكبر في المهمتين. علاوة على ذلك، في الحياة الواقعية، تعتبر الأنظمة المستقلة عن النص أكثر اهتماما من الناحية التجارية من الأنظمة المعتمدة على النص لأنه من الصعب تقليد عبارة غير معروفة أكثر من تلك المعروفة.في هذه الدراسة تم بناء النظام من نوع معتمد على النص.

4. قواعد البيانات

تشكل قواعد بيانات الصوتية الركن الرئيسي في بناء نظم حاسوبية، وتشكل البنية التحتية لبناء نظم تخاطب مع الحاسوب. في هذه الدراسة تم الاعتماد على اثنين من قواعد بيانات الاولى (audioMNIST_meta) تم الحصول عليهل من موقع (Kaggle) ،حيث تم تكوين قاعدة البيانات من قبل مجموعة من الباحثين[13] . اما قاعدة البيانات الثانية تم تكوينها من قبل الباحث (QS-Dataset). تتكون قاعدة البيانات (QS-Dataset) من متحدثين باللغة العربية اما قاعدة البيانات (audioMNIST_meta) فتضم متحدثين باللغة الإنكليزية، وذلك من اجل شمولية عمل النظام على جميع الغات.

تتألف قاعدة البيانات (QS-Dataset) من ملفات صوتية احادية القناة لـ 35 شخص، تضم 25 مقطع صوتي لكل متحدث بالامتداد (.wav)وبمعدل عينة 22050. خلال عملية التسجيل يعمل كل متكلم على تكرار لجملة معينة بمقدار 25 مرة في تسجيل صوتي واحد، ومن ثم خزن التسجيل الصوتي التابع لمتكلم معين في ملف مستقل عن بقية المتكلمين. الجملة التي تم اختيارها هي (سبحان الله وبحمده سبحان الله العظيم) وذلك لضمان الحصول على فترة زمنية أطول.

قاعدة البيانات (audioMNIST_meta) تتألف من 35 شخص، لكل شخص 25 مقطع صوتي. الكلمة التي تم اختيارها هي (One) في كل مقطع يتم نطقها ثلاثة او أربع مرات لضمان حصول على مدة زمينه كافية للتساوي مع قاعدة البيانات الأولى.

5. معالجة الإشارة الرقمية

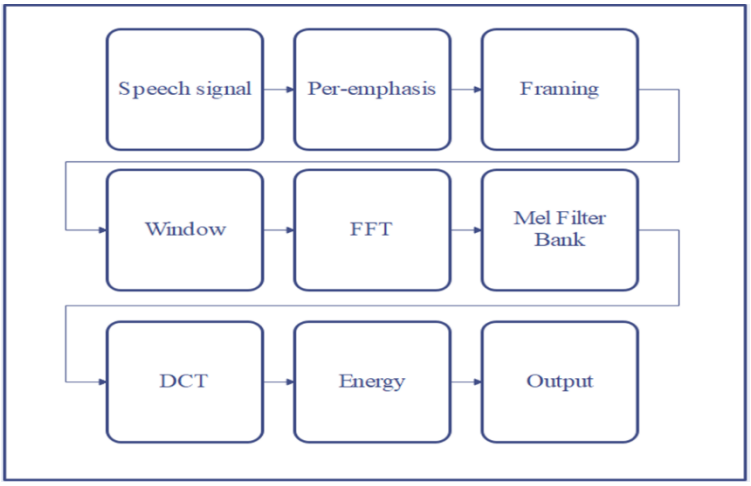
تعرف الاشارة على انها كمية فيزيائية (مادية) متغيرة بتغير الزمان والمكان بالإضافة الى بعض المتغيرات المستقلة الاخرى، اما معالجة الاشارة فهي عملية استخراج معلومات مهمة من الاشارة بطريقة فعالة وقوية. ومن الاشارات المهمة في التواصل بين البشر هي اشارة الكلام، فعندما يتكلم الانسان فإنه ينشئ موجة تناظرية في الهواء تحول عن طريق اللاقط الصوتي الى اشارة تناظرية (Analog) signal، بينما تحول الى بيانات رقمية عندما تخزن على جهاز الحاسوب بشكل (0,1)[14]. ان أحدى الخطوات الرئيسية في معالجة وتحليل الاشارات الصوتية تتمثل بتحويل الاشارة الخام الى تمثيلات تناسب الهدف المطلوبة انجازه، يتم تمييز موجة صوتية عن اخرى من خلال بعض الخصائص منها التردد (Frequency) والسعة (Amplitude) والطور (Phase)[13].

1.5 استخراج الخصائص الصوتية

يمكن القول بأن مرحلة استخلاص الخصائص هي من أهم مراحل التعرف على الكلام؛ لما لها من تأثير على نسبة التعرف الكلية للشبكة العصبية الاصطناعية. نورد فيما يلي واحدة من أهم الخوارزميات المستخدمة في استخلاص الخصائص والتي تسمى خوارومية درجة النغم (MFCC).

1.1.5 خوارزمية تقنية درجة النغم(Mel Frequency Cepstral Coefficients (MFCC

تعتبر خوارزمية MFCC من أهم خوارزميات استخلاص الخصائص وذلك بسبب النتائج الجيدة التي تعطيها. يوضح الشكل(2) مخطط يمثل الخطوات الاساسية لخوارزمية MFCC ، اذ تبدأ بالمعالجة الأولية لإشارة الكلام، ثم تقطيع الإشارة الى اطارات قصيرة الزمن، يلي ذلك عملية الضرب بالنافذة ثم تطبيق تحويل فورير على كل إطار، بعدها يتم تحويل الترددات الى ترددات Mel ومن ثم تجمع الطاقة ضمن حزمة مرشحات، أخيرا تحويل استخدام تحويل جيب التمام المتقطع Discrete Cosine Transform (DCT) للحصول على المعاملات.



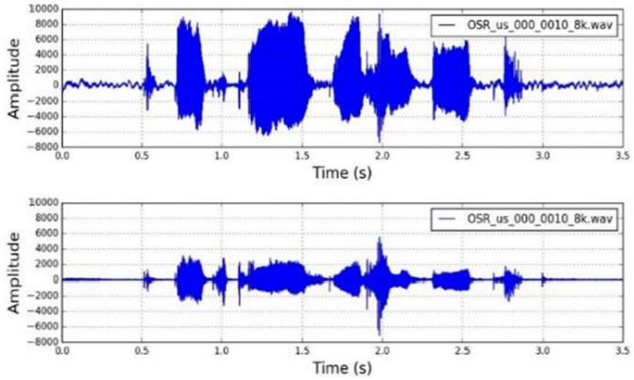
الشكل (2): مخطط لمراحل MFCC[1]

1.1.1.5 التشديد المسبق Pre-emphasis

تتمثل الخطوة الاولى من استخلاص خصائص الاشارة بواسطة MFCC بعملية Pre-emphasis وهي عملية مرشح تردد عالي high pass filter على الإشارة وذلك من أجل تعويض جزء التردد العالي الذي افتقده أثناء آلية إنتاج الكلام (زيادة الطاقة النسبية للطيف عالي التردد) كما موضح في الشكل (3)، حيث يتم إعادة تقييم كل قيمة في إشارة الكلام باستخدام المعادلة(1) التالية:

S2(n) = s(n) – a*s(n-1)... .... (1)

حيث تمثل s(n) إشارة الكلام،s2(n) الاشارة الناتجة من عملية ال pre-emphasis، اما a فهي قيمة ثابتة تتراوح بين (0.9 و 0.1).

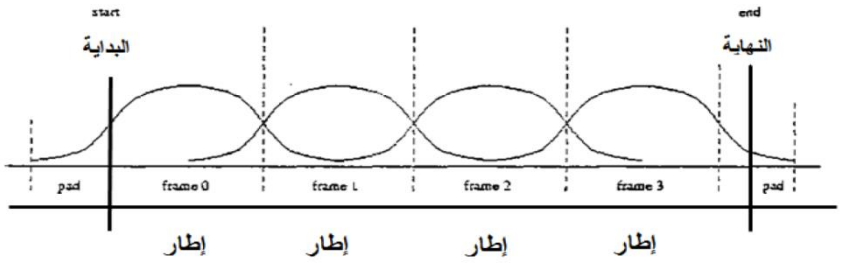


الشكل (3): الإشارة قبل وبعد عمليةPre−emphasis[1]

2.1.1.5 التأطير و النوافذ

نظرا لكون الإشارة الصوتية متغيرة باستمرار في الكلام. يتم نمذجة هذه الإشارة المتغيرة عن طريق إنشاء مقاطع صغيرة مأخوذة من الصوت على أنها ثابتة، وذلك من خلال عملية التأطير وهي عملية فصل العينات من الصوت الخام إلى مقاطع ذات طول ثابت N يشار إليها بالإطارات (Frames) وعادة ما يكون طول الاطار 20ms [1].

بعد تهيئة الإطارات يتم تمرير كل اطار من خلال نافذة بفترة معينة وذلك لتقليل عدم استمرارية الإشارة في بداية ونهاية كل اطار(تقليل انقطاع إشارة الكلام قبل وبعد كل إطار) كما موضح في الشكل (4) [15]. اذا عرفنا النافذة بـ W(n) والاشارة الداخلة بـ $X_1(n)$ فان الإشارة الخارجة هي :



الشكل (4) نموذج لعملية الإطارات والنوافذ

$$X_1'(n) = X_1(n) * W(n) \qquad 0 \le n \le N-1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

هناك العديد من أنواع النوافذ مثل: نافذة هامنغ (Hamming window) ونافذة هاننغ (Hanning window) ونافذة (Blackman window) والنافذة المستطيلة (Rectangular window) والنافذة المثلثة (Triangular window) ونافذة القيصر (Kaiser window) ونافذة غاوس (Gaussian Window) ونافذة ويلش (Welch Window).

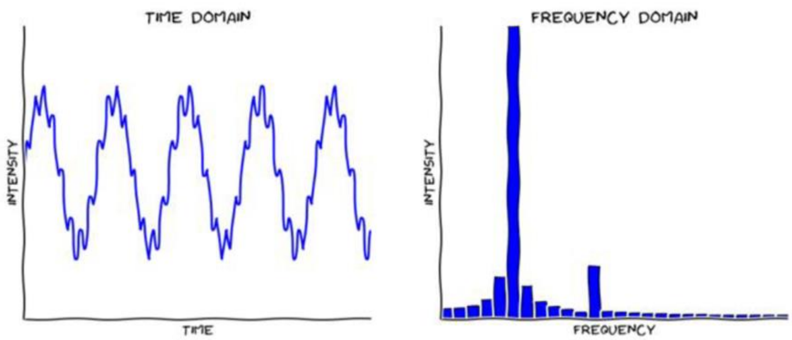
النافذة الأكثر شيوعا في أنظمة التعرف على هوية المتحدث هيا نافذة هامينغ (Hamming window) والمعروفة بالمعادلة (3):

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1 \dots \dots \dots (3)$$

حيث تمثل w(n) العينة الجديدة، n هي تسلسل العينة المفردة و N هو الطول الكلي النافذ.

3.1.1.5 تحويل فورير السريع Fast Fourier Transform (FFT)

بعد عملية الضرب بالنافذة، يتم تطبيق خوارزمية تحويل فورير السريع لكل إطار، وذلك لاستخراج مركبات التردد للإشارة في مجال الزمن [1].كما موضح في الشكل (5).



الشكل (5) يوصف التحويل فورير السريع (FFT)[1]

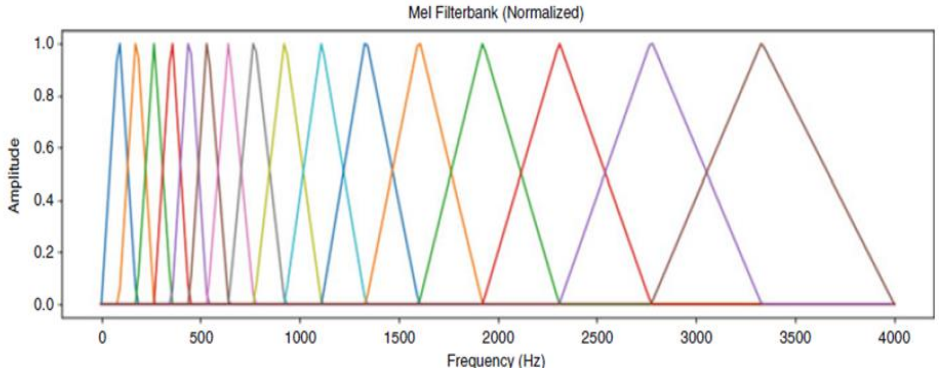
4.1.1.5 بنك المرشحات الميل Mel Filter Bank

وهي عبارة عن مجموعة من مرشحات التي تحاكي النظام السمعي البشري. بدلاً من اتباع مقياس خطي، تعمل هذه المرشحات المثلثية اللوغاريتمية عند الترددات الأعلى والخطية عند الترددات المنخفضة، وهو أمر نموذجي في إشارات الكلام. عادة ما يحتوي بنك المرشح على 40 مرشحًا وكما موضح بالشكل (6). يمكن أن يتم التحويل بين نطاقات Mel (m) و Hertz (f) من خلال:

$$\boldsymbol{m = 2595\log_{10}\left(1 + \frac{f}{700}\right) \dots \dots \dots \dots \dots (4)}$$

$$\boldsymbol{f = 700\left(100^{\frac{m}{2595}} - 1\right) \dots \dots \dots \dots \dots (5)}$$

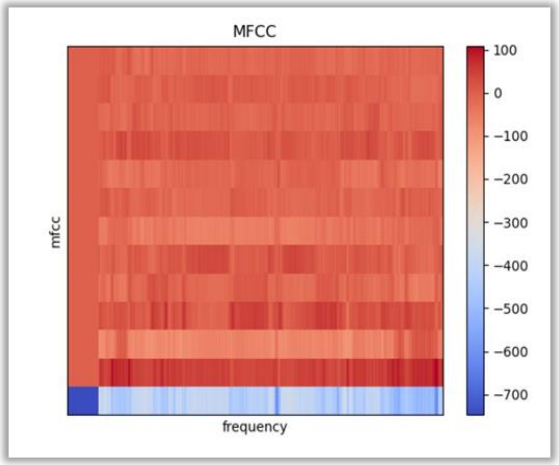
يمثل ناتج عملية الترشيح المجموع المرجح للترددات الطيفية التي تتوافق مع كل مرشح. تحدد هذه القيم ترددات الإدخال في مقياس ميل[16].



الشكل (6) بنك مرشح ميل[17]

5.1.1.5 تحويل جيب التمام المنفصل Discrete Cosine Transform (DCT)

يقوم تحويل جيب التمام المنفصل (DCT) بتعيين ميزات مقياس Mel إلى مجال الوقت. الدالة DCT مشابهة لتحويل فورير ولكن يستخدم أرقام حقيقية فقط (تحويل فورير تنتج أرقام معقدة). وهو يضغط بيانات الإدخال في مجموعة من معاملات جيب التمام التي تصف التذبذبات في الدالة. ويشار إلى الناتج من هذا التحويل باسم MFCC كما موضح في الشكل (7). بعد استخلاص الخصائص يتم خزن النتائج بشكل (Vector) في ملف اسمه Code_ Dictionary.



الشكل (7): التمثيل الطيفي لمعاملات درجة النغم

6. ملف القاموس Code_Dictionary

عبارة عن ملف يتم فيه خزن النتائج بعد عملية استخلاص الخصائص حيث يتكون هذا الملف من ثلاثة حقول, الحقل الاول يضم أسماء المتحدثين والحقل الثاني يضم ارقام تمثل (labels) والذي يمثل فهرسة لهذه الاسماء المتحدثين والحقل الثالث يضم الخصائص لكل متحدث.يوضح الشكل(8) هيكلة الملف القاموس.

```
{
    "mapping": [
        "",..........,""
    ],
    "index": [
        "",..........,""
    ],
    "mfcc": [
       [],............[],
    ]
}
```

الشكل(8): هيكلة الملف القاموس

7. الشبكات العصبية الاصطناعية

تعرف الشبكات العصبية الاصطناعية (Artificial Neural Networks (ANN))على انها نماذج حسابية ذات قدرة على تجميع وتنظيم البيانات وتعلم المعلومات المعممة من خلال الاعتماد على المعالجة المتوازية. تتكون الشبكة العصبية من مجموعة من الطبقات (Layers) مكونة من وحدات معالجة تسمى بالخلايا (Neurons) متماثلة تتصل فيما بينها عن طريق ارسال الاشارات الموزونة الى بعضها البعض، يتم في مرحلة التعلم تحديث المعاملات لتحقيق شرط معين وذلك عن طريق استخدام مجموعة التدريب (Learning Set) التي تستخدم لحساب نسبة الخطأ [17].

1.7 الشبكة العصبية متعددة الطبقات

تتألف هذه شبكة العصبية من طبقة واحدة او عدة طبقات من الطبقات المخفية. ان فائدة الطبقات المخفية هو اكتشاف المزايا (features) الموجودة في الإشارات الداخلة اليها [14]. توجد العديد من الطرق لتدريب الشبكات ومن أشهرها هي الانتشار الامامي(forward-propagation) والانتشار العكسي(back-propagation). يتم الادخال المدخلات الى طبقة الادخال تمر الى طبقات المخفية ويتم المعالجة هذه الادخالات تما تمرر وتعالج مرة أخرى طبقة الإخراج تسمى هذه خوارزمية التدريب الانتشار الامام (forward-propagation). اما التدريب الشبكة عن طريق الانتشار العكس(back-propagation) حيث يتم تقديم انماط الادخال للتدريب إلى طبقة الادخال الشبكة. تقوم الشبكة بعد ذلك بنشر نمط الإدخال من طبقة إلى طبقة حتى يتم إنشاء نمط الإخراج بواسطة طبقة الإخراج. يتم حساب الخطأ في حال ظهور النتائج غير مرغوبة ثم يتم نشره للخلف من طبقة الإخراج إلى طبقة الإدخال وتستمر العملية لحين الحصول على النتائج المطلوبة [17].

8.الشبكات العصبية العميقة

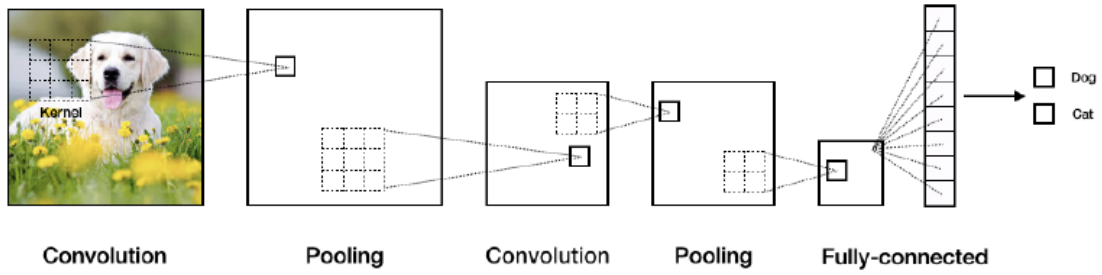
أصبح التعلم العميق طريقة مثيرة للاهتمام وقوية لتعلم الآلة مع تطبيقات ناجحة في العديد من المجالات ، مثل معالجة اللغة الطبيعية ، والتعرف على الصور ، والتعرف على الأحرف المكتوبة بخط اليد، والتعرف على المتحدث ، ورؤية الحاسوب، حيث أظهر التعلم العميق نجاحًا في التعرف على الكلام وتحديد هوية المتحدث على الطرق التقليدية مثل تلك التي تستخدم معاملات تردد ميل للتعرف على المتحدث باستخدام نماذج خليط غاوسي .

ان التعلم العميق يوفر طريقة اكثر تكيفا من خلال استخدام الشبكات العصبية العميقة التي تتعلم الخصائص من بيانات الادخال وبالتالي فإنها تجعل الحاسوب قادر على اتخاذ القرار، تعتمد الطرق التعلم العميق على الطرق التي تستند الى تعلم تمثيل البيانات .

8-1 الشبكة العصبية الالتفافية

تأتي تسمية هذا النوع من الطبقات من عملية الطي أو الالتفاف الرياضية حيث تطبق في طبقات التفافية مُرشِّحًا (Filter ويُعرف أيضًا بـ kernel) بعدد N يحدد حسب النتائج اثناء عملية التدريب، ويمكن تمثيلها بشكل مصفوفة، من شأنه تحديد وجود سمات أو أنماط مُعينة في الصورة الأصلية. يكون حجم المُرشِّح صغير ليمسح مصفوفة الادخال بشكل كامل ويطبق العمليات الحسابية بغية استخراج السمات (Features). يُعاد ضبط قيم المُرشِّح خلال عملية التدريب الدورية، وعند تدريب الشبكة، توظَّف الطبقات المخفية الأولى في استخراج السمات البسيطة والواضحة، مثل الحواف في الاتجاهات المُختلفة، ومع التعمّق أكثر في الطبقات المخفية في الشبكة، تزداد درجة تعقيد السمات التي يجب تحديدها واستخراجها[18].الشكل (9) معمارية الشبكة العصبية الالتفافية (CNN).

الشكل (9): نموذج الشبكة العصبية الالتفافية (CNN) [19].



9. بناء النظام

في هذه الدراسة تم الاعتماد على الشبكات العصبية (الشبكة العصبية الالتفافية والشبكة العصبية متعددة الطبقات) لبناء نظام التعرف على المتحدث. حيث تمت بناء هيكلية الشبكات بناءً على التجربة.

1.9 هيكلية الشبكة العصبية متعددة الطبقات

تتألف هيكلية الشبكة من 4 طبقات مرتبطة ارتباطا كليا ، بالإضافة الى طبقة الادخال والتي تحتوي على 1131 خلية و3 طبقات مخفية، تحتوي الطبقة الأولى على 256 خلية اما الطبقة المخفية الثانية فتحتوي على 256 خلية بينما الطبقة المخفية الثالثة فتحتوي على 128 خلية، تم استخدام دالة التنشيط relu في الطبقات المخفية لتدريب الشبكة وتحسين الاوزان، اما طبقة الإخراج فتتألف من 36 خلية وهي تمثل هذه الطبقة عدد المتحدثين، اما دالة التنشيط في طبقة الاخراج فهي Softmax. يوضح الجدول (1) معمارية الشبكة MNN.

الجدول (1): معمارية الشبكة MNN

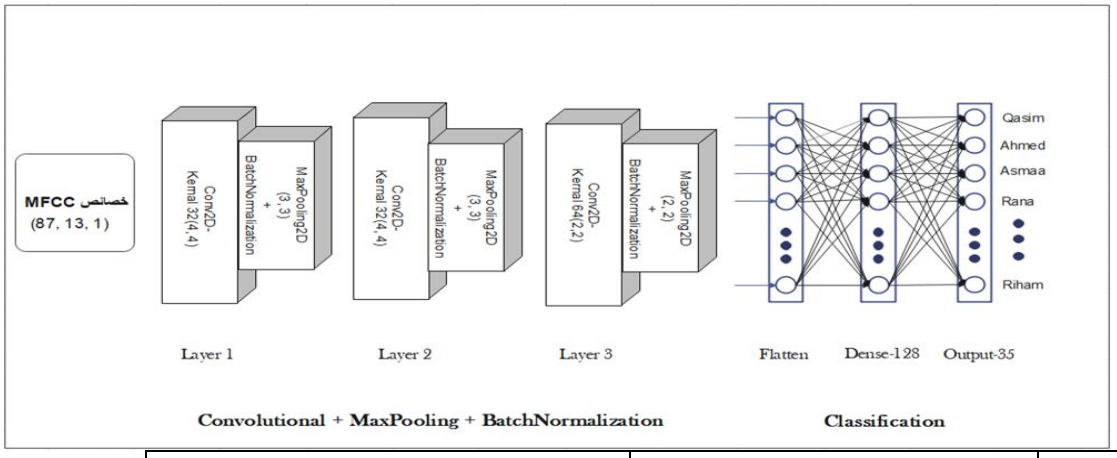
Layer (type)	Output Shape	Param #
Flatten (Flatten)	(None , 1131)	0
Dense (Dense)	(None , 128)	144896
Dense_1 (Dense)	(None , 128)	16512

Dense_2 (Dense)	**(None , 128)**	**16512**
Dense_3 (Dense)	**(None , 35)**	**4515**

2.9 الشبكة العصبية الالتفافية

تتألف شبكة CNN من 5 طبقات ، تتكون الطبقة الأولى من طبقة التفافية ثنائية (Conv2D) ، يتم في هذه الطبقة استخدام (64) فلتر (Kernel)كل فلتر يتمثل بمصفوفة ثنائية بحجم (4*4) و دالة تنشيط نوع relu ،بالإضافة الى الطبقة الالتفافية ،تم استخدام التجميع (MaxPooling2D) بحجم (3,3) والتي تعمل على تبسيط المعلومات التي تم الحصول عليها بواسطة الطبقة الالتفافية والخطوة (strides) بحجم (2,2) لتحديد موقع الفلتر في المرحلة التالية بالنسبة للمرحلة الحالية ، ونوع الحشو (padding) التي تعمل على اضافة اصفار الى البيانات المدخلة و في الاتجاهات الاربع ، تهدف هذه العملية الى جعل ابعاد الاخراج مساوية لإبعاد الادخال، ولزيادة سرعة عملية التدريب يتم استخدام دفعة التطبيع (Batch Normalization) ،اما الطبقة الثانية فهي شبيهة بالطبقة الالتفافية الأولى مع اختلاف حجم مصفوفة الفلتر الى (3*3)، بينما تضم الطبقة الثالثة 64 فلتر بحجم (2,2) ودالة تنشيط نوع relu ، بالإضافة الى MaxPooling2D بحجم (2, 2) strides بحجم (2, 2) ونوع الحشو صفري ايضا. الطبقة الرابعة هي طبقة تسطح flatten تتكون من 640 خلية، وطبقة تكيف (Dense) تضم 128 خلية والطبقة الاخيرة مكونة من 35 خلية وبدالة تنشيط نوع دالة السينية softmax. الشكل يوضح الجدول (2) معمارية الشبكةCNN.كما يوضح الشكل (10)هيكلية الشبكة العصبية الالتفافية (CNN) .

الشكل (10): يوضح هكيلية الشبكة العصبية الالتفافية (CNN)



الجدول (2): معمارية الشبكةCNN

Layer (type)	Output Shape	Param #
Conv2d (conv2D)	(None , 84 , 10 ,92)	1564
Max_pooling2d (maxpooling2D)	(None , 42 , 5 ,92)	0
Batch_normalization	(None , 42 , 5 ,92)	368
Conv2d_1(conv2D)	(None , 42 , 5 ,92)	76268
Max_pooling2d_1	(None , 20 , 2 ,92)	0
Batch_normalization_1	(None , 20 , 2 ,92)	368
Conv2d_2(conv2D)	(None , 19 , 1 ,92)	33948
Max_pooling2d_2	(None , 10 , 1 ,92)	0

Batch_normalization_2	(None , 10 , 1 ,92)	368
Flatten_1	(None,920)	0
Dense_4	(None,920)	117888
Dense_5	(None,920)	4515

3.9 تقسيم قاعد البيانات

بعد الانتهاء من اعداد هيكلية الشبكات بشكل كامل، لابد من تقسيم قاعدة البيانات الى أجزاء، قسم منها للتدريب واخر للاختبار وقسم اخر لتدقيق صحة النتائج التي تستخدم فقط مع شبكة CNN. تم تقسيم قواعد البيانات الصوتية حسب الجدول (3).

الجدول (3): يوضح الية التقسم قاعدة البيانات الصوتي

الشبكة	التدريب	التحقيق	الاختبار
MNN	%70	–	%30
CNN	%40	%30	%30

4.9 مرحلة التدريب

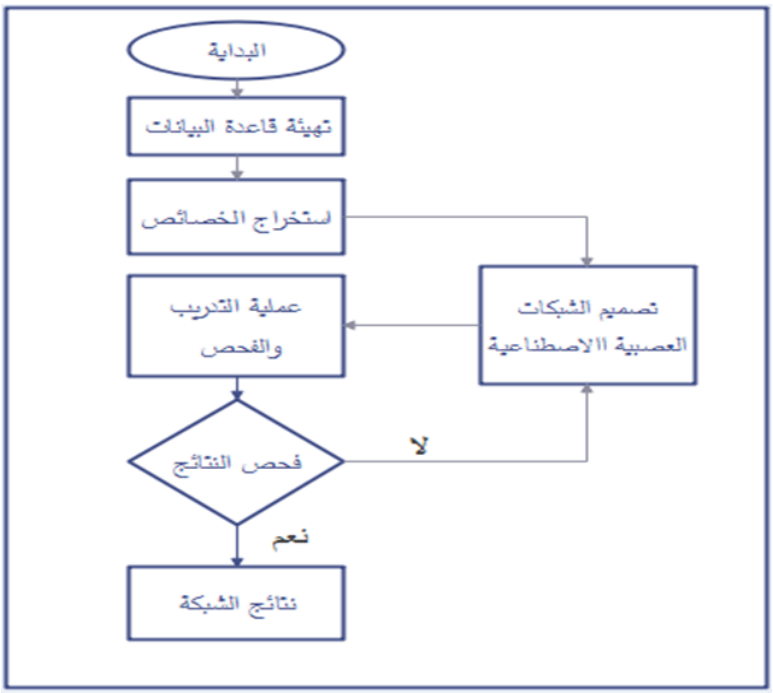
تشكل مرحلة تدريب الشبكة العصبية مرحلة أساسية وهامة تأتي بعد مرحلة اعداد هيكلية الشبكة وتقسيم البيانات، حيث يجري تدريبها على مجموعة من الأمثلة لتعطي الشبكة النتائج صحيحة على كل الأمثلة التي تدربت عليها. وقبل التدريب لابد من تهيئة معدل التعليم (Learning rate) والتي تكون قيمتها (0.0001)، وكذلك تهيئة مراحل التدريب (Epoch) والتي تكون 30 دورة تدريب وبعدد الأمثلة (batch size) والتي تكون 10.

5.9 مرحلة الاختبار

اختبار الشبكة مشابه تماماً لعملية التدريب إلا أن الشبكة في هذه المرحلة لا تضبط أوزانها، وإنما فقط تقوم بعمليتي الجمع والتحويل ومقارنة الناتج الذي تنتجه الشبكة بالناتج الهدف. حيث يتم عرض فئة اختبار على الشبكة وتحتوي هذه الفئة على مجموعة من المدخلات والمخرجات المصاحبة لكل مدخل. وبعد اكتمال مرحلة التدريب والاختبار.

6.9.الية عمل النظام

في البداية يتم تهيئة قاعدة البيانات ثم يتم استخلاص الخصائص وخزن النتائج في ملف القاموس **Code_Dictionary**، يتم اخذ الخصائص من القاموس والتي تعتبر كمدخلات للشبكات العصبية, بعد ذلك يتم تدريب الشبكة وضبط معاملات الشبكة واخيرا يتم اختبار الشبكة والحصول على النتائج الدقة (Accuracy)، وبعد الحصول على النتائج يتم التأكد من صحة البرنامج عبر مقاييس (FAR,FRR).الشكل (11) يوضح مخطط العام للنظام.



الشكل (11): يوضح مخطط العام للنظام

10- مقاييس النتائج والتحقق منها

المقاييس التي نعتمد عليها في هذه الدراسة في تنبؤ الشبكة (Accuracy)، واما التحقق من صحة النظام عبر مقاييس (FAR، FRR)، معدل القبول الخاطئ (FAR)، هي النسبة المئوية لحالات تحديد الهوية التي يتم فيها قبول الأشخاص غير المصرح لهم بشكل غير صحيح . اما معدل الرفض الخاطئ (FRR)، النسبة المئوية لحالات تحديد الهوية التي يتم فيها رفض الأشخاص المصرح لهم بشكل غير صحيح. توضح المعادلات التالية كلما ذكر في هذه الفقرة [17]:

$$Accuracy = \frac{No.of\ correct\ recognitions}{Totsl\ no.of\ trials} \dots\dots\dots\dots\dots\dots\dots\dots (6)$$

$$FRR = \frac{No.of\ true - speakers\ rejected}{Total\ no.of\ true - speaker\ trials} \dots\dots\dots\dots\dots\dots\dots\dots (7)$$

$$FAR = \frac{No.of\ impostors\ accepted}{Total\ no.of\ impostor\ attempts} \dots\dots\dots\dots\dots\dots\dots\dots\dots (8)$$

11. النتائج والمناقشة

نتيجة للعملية السابقة هي تحويل البيانات الصوتية إلى العديد من ميزات الخصائص، تعتبر هذه الخصائص كمدخلات للشبكات العصبية، في الفقرات التالية سيتم مناقشة النتائج النظام التي تما الحصول عليها باستخدام الشبكة العصبية الالتفافية (CNN) والشبكة الصبية متعددة الطبقات (MNN).

1.11 النتائج التطبيقية

تختلف النتائج لعملية التعرف على المتحدث باختلاف الشبكات العصبية وجودة قاعدة البيانات الصوتية المستخدمة لعملية التعرف على المتحدث، لذا فقد تما قياس كفاءة اداء الشبكات العصبية باستخدام مقياس الدقة (Accuracy)، وعلمية التحقق عن طريق مقاييس (FARوFRR). فيما يلي الجداول النتائج الدقة التي تما الحصول عليها من الشبكة (CNN) وشبكة (MNN) والتي تعتمد على عدد المتحدثين:

الجدول (4): النتائج الدقة للشبكات العصبية بناء على قاعدة بيانات(QS-Dataset)

No. of Speakers	Accuracy	
	MNN	CNN
5	%100	%70
10	%88	%89
15	%84	%95
20	%85	%96
25	%82	%97
30	%81	%96
35	%80	%96

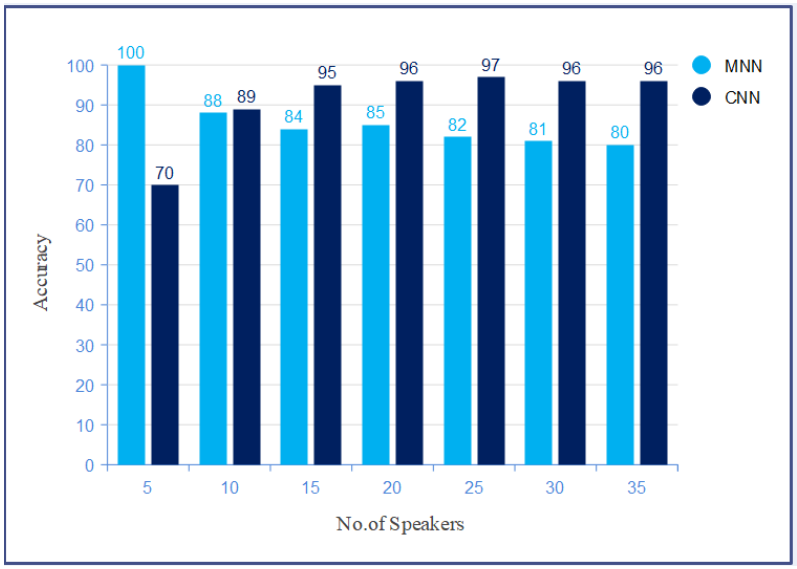
الجدول (5):النتائج الدقة للشبكات العصبية بناء على قاعدة بيانات (audioMNIST_meta)

No. of Speakers	Accuracy	
	MNN	CNN
5	%87	%53
10	%76	%79
15	%70	%83
20	%66	%81
25	%63	%82
30	%56	%82
35	%56	%83

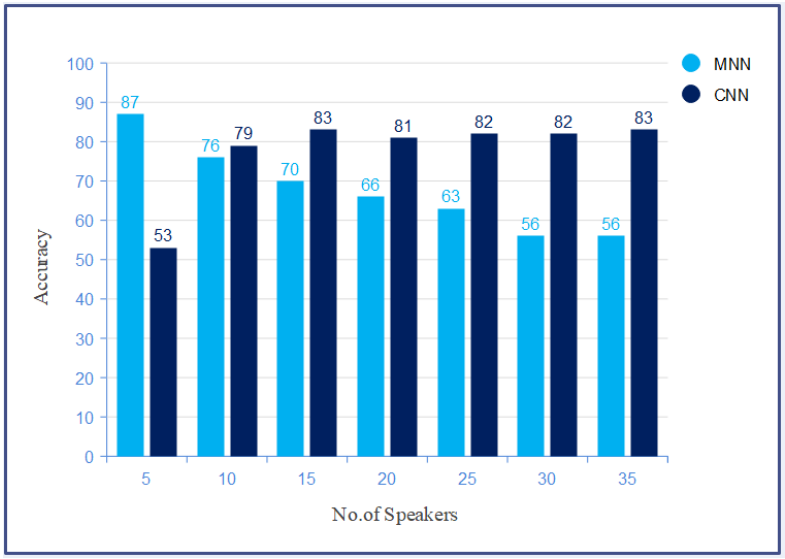
1.1.11 مناقشة النتائج

في الجدول (4) تحتوي على نتائج الدقة (Accuracy) لعملية التعرف على هوية المتحدث باستخدام الشبكة العصبية متعددة الطبقات (MNN) والشبكة العصبية الالتفافية (CNN) بالاعتماد على قاعدة البيانات الصوتية (QS-Dataset)، نلاحظ النتائج شبكة (MNN) عالية جدا بنسبة تصل الى 100% عندما يكون عدد المتحدثين 5 وعند يرتفع عدد المتحدثين تقل هذه نسبة تدريجيا الى ان تستقر هذه النسبة لدرجة معينة وذلك لأنه نوعية هذه الشبكة تتعامل مع بيانات القليلة بالمقارنة مع الشبكة العصبية الالتفافية (CNN)

تكون عكس النتائج الشبكة (MNN) حيث تكون الشبكة ذات كفاءة قليلة كلما كانت عدد البيانات قليلة بعكس ما اذا توفرت بيانات كبيرة تكون نتائج الشبكة عالية جدا ، اما النتائج الدقة في الجدول (5) تكون مختلفة عن النتائج الجدول (4) يعود ذلك الى السبب الرئيسي هو جودة الإشارة الصوتية. جميع النسب الموجودة في جميع الجداول التي تعود للفقرة (1.5) هيا نسب تقريبية أي تكون اعلى او اقل ب 1% الى 5% والسبب في ذلك يعود الى القيم الأولية التي تعطى للمعاملات الشبكة بشكل تلقائي عند بداية التدريب ولكون هذه الشبكات محددة بعدد فترات تدريبية ، حيث عند كل مرحلة تدريب تعطي قيماً مختلفة للمعاملات ولضمان حصول على نسبة تقريبية جيدة يتم تدريب الشبكات لـ 5 مرات وفي كل مرة يتم تسجيل النسبة التي تما الحصول عليها من شبكات وبعد ذلك يتم جمع جميع القيم وتقسيمها على 5 للحصول على النسبة المتوقعة. الشكل (12) و(13) يبين النتائج الشبكات بشكل مخطط بياني .



الشكل (12) مخطط بياني لنتائج الشبكات المبنية على قاعدة بيانات (QA-Dataset)



الشكل (13) مخطط بياني لنتائج الشبكات المبنية على قاعدة بيانات (audioMNIST_meta)

2.11 نتائج التحقق من النظام

بعد الحصول على النتائج الدقة (Accuracy)، لابد من التحقق من عمل النظام عن طريق مقاييس (FAR وFRR)، تم اخذ عينات صوتية من القواعد البيانات والبالغ عددها 875 مقطع صوتي،يتم اخبار النظام ان جميع هذه الأصوات تعود لمتحدث معين ،يقوم النظام بالتحقق من جميع مقاطع الصوتية وبعدها تخزن الناتج ثم تكرر العملية على بقية المتحدثين والحصول على ناتج نهائي ،تسمى هذه العملية بـ معدل القبول الخاطئ (FAR).اما مقياس معدل الرفض الخاطئ (FRR)، فيؤخذ عينات صوتية تعود للمتحدث نفسه ويقوم النظام بفحص نسبة فشل النظام في التعرف على المتحدث نفسة.فيما يلي النتائج التي تم التوصل اليها :

الجدول (6): نتائج التحقق من الشبكة العصبية متعددة الطبقات (MNN)

Dataset	FAR	FRR
QS-Dataset	%2.85	%10.4
audioMNIST_meta	%2.85	%24.34

الجدول (7): نتائج التحقق من الشبكة العصبية الالتفافية (CNN)

Dataset	FAR	FRR
QS-Dataset	%2.85	%3.88
audioMNIST_meta	%2.85	%11.31

1.2.11 مناقشة النتائج التحقق من النظام

في الجدول (6) و (7) نلاحظ ان نسبة مقياس (FAR)هيا 2.85% ،وسبب في ذلك عند فحص النظام وفق مقاطع الصوتية المخصصة لهذه الغرض كانت من ضمن هذه مقاطع صوتية أصوات تعود نفس الأشخاص المسجلين في النظام .اما القيم التي تعود لمقياس (FRR)،نلاحظ في الجدول (6) قيم اعلى مما هيا موجوده في الجدول (7) وسبب في ذلك يعود الى ان الأنظمة التي بنيت بالاعتماد على شبكة MNN تكون اقل دقة واقل فعالية عندما تتعامل مع البيانات الكبيرة بعكس الأنظمة التي تم بناءها بالاعتماد على شبكة CNN تكون اكثر كفاءه واكثر دقة .

12.الاستنتاجات

تقدم هذه الدراسة أحدث انظمة التعرف على المتحدثين والتحقق منهم، حيث اعتمد على أساليب التعلم العميق التي تتضمن الشبكات العصبية (MNN وCNN) للتعرف على المتحدثين وكذلك مقاييس التحقق من النظام (FAR وFRR). تم اختبار النظام على مجموعة من قواعد البيانات (QS−Dataset ، audioMNIST_meta) ، حيث يوضح نسبة الدقة التي حصلت عليها عند استخدام قاعدة البيانات (QS−Dataset) وكذلك نسبة الدقة التي حصلت عليها عند استخدام قاعدة البيانات (audioMNIST_meta)،ويمكن

ملاحظة عند استخدام اقل عدد ممكن من البيانات تكون شبكة MNN اكثر فعالية بينما اذا كانت كمية البيانات كبيرة تكون شبكة CNN اكثر فعالية. يتميز الأنظمة التي تما بناءها بسرعة التدريب الشبكات وباقل وقت وبعدد قليل من عينات التدريب وكذلك ما يميز هذه الأنظمة انه لا تحتاج الى حاسبات ذات مواصفات عالية في عملية تدريب الشبكة .

شكر وتقدير

يتقدم الباحثون بالشكر للمراجعين والمحررين على ملاحظاتهم المفيدة. كما يتقدم الباحثون بالشكر والتقدير لقسم علوم الحاسوب، جامعة الموصل على تقديم الدعم والتسهيلات اللازمة لإكمال هذا البحث.

13. المصادر

[1] B. Alkhatib and M. M. K. Eddin, "Voice Identification Using MFCC and Vector Quantization," *Baghdad Sci. J.*, vol. 17, no. 3 2020 ,ملحق.

[2] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," in *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*, 2012, pp. 248–251.

[3] S. Bunrit, T. Inkian, N. Kerdprasop, and K. Kerdprasop, "Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 2, pp. 143–148, 2019.

[4] S. V Ault, R. J. Perez, C. A. Kimble, and J. Wang, "On speech recognition algorithms," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 6, pp. 518–523, 2018.

[5] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[6] I. GAVAT and D. MILITARU, "New trends in machine learning for speech recognition," in Annual Symposium of the Institute of Solid Mechanics and Session of the Commission of Acoustics SISOM, At Bucharest, Romania, 2015, vol. 2015, pp. 271–276.

[7] P. Sharma and A. Garg, "Feature Extraction and Recognition of Hindi Spoken Words using Neural Networks," Int. J. Comput. Appl., vol. 142, no. 7, pp. 12–17, 2016.

[8] A. Antony and R. Gopikakumari, "Speaker identification based on combination of MFCC and UMRT based features," Procedia Comput. Sci., vol. 143, pp. 250–257, 2018.

[9] V. Thanda Setty, "Speaker Recognition using Deep Neural Networks with reduced Complexity," 2018, M.sc. Thesis, Texas State University.

[10] S. Safavi, P. Jancovic, M. J. Russell, and M. J. Carey, "Identification of gender from children's speech by computers and humans.," in INTERSPEECH, 2013, pp. 2440–2444.

[11] R. Jagiasi, S. Ghosalkar, P. Kulal, and A. Bharambe, "CNN based speaker recognition in language and text-independent small scale system," in 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2019, pp. 176–179.

[12] S. Abd El-Moneim, M. A. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. Abd El-Samie, "Text-independent speaker recognition using LSTM-RNN and speech enhancement," Multimed. Tools Appl., vol. 79, no. 33, pp. 24013–24028, 2020.

[13] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," arXiv Prepr. arXiv1807.03418, 2018.
[14] R. A. K. Nilu Singh, "Digital Signal Processing for Speech Signals," Bilingual international conference of information technology, pp. 134–138.
[15] yusra faysal Ali nesaf, "Distinguishing single-pronounced Arabic numerals using a genetic algorithm," *AL-Rafidain Journal of Computer Sciences and Mathematics*, vol. 1(11), 2013.
[16] Kamath, Uday, John Liu, and James Whitaker. Deep learning for NLP and speech recognition. Vol. 84. Cham: Springer, 2019.
[17] Michael Negnevitsky, "Artificial Intelligence: A Guide to Intelligent Systems.," in *Polyhedron*, Third Edit., vol. 123, University of Tasmania: Addison Wesley, 2005, pp. 165–216.
[18] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights Imaging*, vol. 9, no. 4, pp. 611–629, 2018.
[19] J. Van Der Donckt, "Latent representations for spoken language," 2019, M.sc. Thesis, Gent Universiteit.