الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا
قسم الاتصالات

أعدّت هذه الأطروحة لنيل

درجة الماجستير في معالجة الإشارة

استخدام الشبكات العصبونيّة العميقة للتعرّف على الصوتيمات

Deep Neural Networks for Phonemes Recognition

إعداد

م. محمد سميط

إشراف

د. عبد الناصر العاسمي

30/10/2019

المعهد العالي للعلوم التطبيقية والتكنولوجيا

Higher Institute for Applied Sciences and Technology
Institut Supérieur des Sciences Appliquées et de Technologie

مؤسسة سورية حكومية للتعليم العالي أحدثت في عام 1983، بهدف إعداد أطر متميزة مؤهلة للبحث العلمي والتطوير في مجال العلوم التطبيقية والتقانة، لتساهم بفاعلية في التنمية العلمية والصناعية والاقتصادية في القطر. يشكّل التأهيل الهندسي والدراسات العليا في المعهد العالي محور عمليّة إعداد الأطر المتخصّصة. يخرّج المعهد العالي مهندسين متميزين، بعد دراسة لمدة خمس سنوات، في اختصاصات الاتصالات والمعلوماتية والنظم الإلكترونية والميكاترونيكس وهندسة الطيران وعلوم وهندسة المواد. كما يمنح المعهد العالي درجة الماجستير الأكاديمي، ماجستير بحثي يمتد على سنتين، من خلال مجموعة من برامج الماجستير في نظم الاتصالات وفي التحكم والروبوتيك وفي علوم المواد وفي نظم المعلومات واتخاذ القرار وفي نظم معالجة المعطيات الكبيرة. وأخيراً يمنح المعهد العالي درجة الدكتوراه في عدة اختصاصات موازية لما ذكر في برامج الماجستير. يعتمد المعهد العالي في تميزه على تركيزه على النوع وليس الكم، فهو ينتقي في المرحلة الهندسية شريحة الطلاب المتفوقين في شهادة الدراسة الثانوية السورية من الفرع العلمي أو من في حكمهم. أما في مرحلة الماجستير فيقبل المعهد العالي حملة الشهادات الجامعية المواتية للماجستير المطلوب، وذلك على أساس مفاضلة خاصة لاختيار الأفضل. كما يقدم المعهد للطلاب جواً متميزاً للدراسة والبحث بدءاً من كوادره المتفرغة عالية التأهيل ومناهجه المواكبة للتطورات العلمية، وانتهاءً بإمكانيات مختبراته المتميزة في القطر وبنيته التحتية المواتية من صالات حواسب وورش ومقدرات مادية وشبكات تعاون مع الصناعة والهيئات الأكاديمية داخل وخارج القطر. كل ذلك في وسائل للراحة والترفيه من سكن طلابي مطعم وصالات رياضية وملاعب ونشاطات طلابية متنوعة. بالإضافة إلى نشاطه الأكاديمي، يضم المعهد العالي أقساماً علمية ومخابر متنوعة ومراكز تكنولوجية، كمخبر الدراسات البيئية ومركز تقانات اللحام ومركز الحوسبة عالية الأداء. تقدم هذه الفعاليات خدمات واستشارات للقطاعين العام والخاص، بالإضافة إلى المشاريع التطويرية والنشاطات البحثية والدورات التدريبية وتنظيم ورش العمل والمؤتمرات العلمية.

الإهداء

إلى تلك الروح التي أنارت لي العالم بأسره، تلك الروح التي كلما ذكرت رفعت رأسي عالياً، تلك الروح التي علمتني الصبر والقوّة، تلك الروح التي أعطتني مفاتيح الحياة، تلك الروح التي ما بخلت وجادت، تلك الروح التي أدفع عمري ثمناً لرجوعها، تلك الروح التي منحتني كل الأمان فكانت دولتي وجيشي وكل حرّاسي، شكراً يا سندي....

والدي المرحوم صلاح الدين سميط 2019/6/14

إلى تلك الحنونة التي وقفت معي في كل لحظاتي، في ضعفي قبل قوتي، في مرضي قبل صحتي، تلك الحنونة التي تخشى عليّ أكثر من نفسي، تلك الحنونة التي جُعلت الجنة تحت قدماها، تلك الحنونة الصابرة، تلك الحنونة التي ربّتني، تلك الحنونة التي لا تجعلني أشعر بمرارة الحياة، شكراً...

أمّي الغالية

إلى تلك الحبيبة التي آمنت بي منذ الصغر، تلك الحبيبة القويّة المجنونة الذكيّة الراقية الحنونة الشغوفة المختلفة، تلك الحبيبة التي كانت وستبقى ملح الحياة وطعم السعادة، تلك الحبيبة المتربّعة على عرش قلبي، تلك الحبيبة الحلم والأمل والماضي والحاضر والمستقبل، تلك الحبيبة التي لن تكفيها كلماتي، شكراً....

رزان

إلى جبالي الرواسي.. ورياتي الشامخة..

إلى من كانوا ومازالوا درعاً في ظهري.. وصخرة أتكئ عليها دون مخافة انزلاقها..

إلى أخوتي: رامي – ريما – مازن – دعاء

إلى تلك البراعم التي أرى الحياة فيها..

شام – صلاح – لين – جاد

إلى أخي وصديق عمري، الغائب الحاضر، الذي لا ينسى أبداً...

الشهيد محمد سميط

إلى أصدقائي الذين كانوا ومازالوا أخوة لي وسند، أصحاب الضحكة والفرح والحزن والألم والحلو والمر والعقل والجنون

عقبة – علي– تمام

إلى كل أهلي وأصدقائي وزملائي في العمل والدراسة، وإلى كل من وضع منارة في طريقي، وإلى كل معلّم علمني حرفاً.

محمد صلاح الدين سميط

“If you want to find the secrets of the universe, think in terms of energy, frequency and vibration.”

Nicola Tesla... (1856-1943)

كلمة شكر

أتقدّم بجزيل الشكر إلى عموم كوادر المعهد العالي للعلوم التطبيقية والتكنولوجيا أساتذةً ومشرفين ومديرين. وأتقدّم بالشكر إلى مشرفي الدكتور عبد الناصر العاسمي على جهوده وملاحظاته القيّمة. كما أتقدّم بالشكر إلى عموم كوادر قسم الرادار وأخص بالذكر الدكتور رضوان قسطنطين والمهندس سليمان جاهوش. وأشكر الدكتور شادي بيطار على كل الدعم لهذا العمل، وأشكر المهندس أحمد مطر على كل ملاحظاته ومساندته، والمهندس محمود عليّان على مساعدته، والعميد المهندس خالد سرغاني على الدعم التقني لهذا العمل، والمهندس عقبة عباس على نقاشاتنا المفيدة، والمهندس وائل رزوق والمهندس محي الدين بندقجي والمهندس سيمون طربوش والمهندس نور الملحم، وأشكر كل من ساعد بحرف أو وضع لمسة في طريقي وطريق هذا العمل...

محمد سميط

النشر العلمي في الأطروحة

تم نشر مقالة بعنوان "**Cascade Deep Neural Network Classifiers for Phonemes Recognition**" ضمن مجلّة محكّمة "**Journal of Engineering and Applied Sciences**"، ومصنّفة من قبل **Scopus** ضمن الربع الثالث. (ستنشر المقالة في العدد الرابع لعام 2020)

Medwell Journals
Scientific Research Publishing Company

Medwell Journals
Tel: +92-41-5003000
Fax: +92-41-8815599
http://medwelljournals.com

September 13, 2019

Dear mohammad smit,

Based on the reviewer's recommendations, I am delighted to inform you that your following manuscript has been accepted for the publication in *Journal of Engineering and Applied Sciences*.

Title	Cascade deep neural networks classifiers for Phonemes recognition
Authors	Mohammad Smit, Abdel-Nasser Al-Assimi
Received on	September 09, 2019
Accepted on	September 13, 2019

Thank you very much for submitting your article to *"Journal of Engineering and Applied Sciences"*.

We look forward to receive more articles in future.

Best Regards

Muhammad Kamran
Journal of Engineering and Applied Sciences

تم أيضاً إرسال مقالة علميّة بعنوان "**التعرّف على الصوتيمات باستخدام الشبكات العصبونية العميقة**" إلى مجلة العلوم الهندسيّة في جامعة دمشق، وتم الحصول على الموافقة النهائية للنشر.

الملخص

أخذت الشبكات العصبونية العميقة في السنوات القليلة الماضية مسألة التعرّف الآلي على الصوت إلى مستويات جديدة من الدقة [1]. حيث حازت على أعلى نسب للتعرّف، سواء على الكلمات بشكل مفرد أو على الصوتيمات. تمثل مسألة التعرّف على الصوتيمات المرحلة الأولى من مراحل التعرّف في أنظمة التعرّف الآلي على الكلام. نقدم في هذا البحث التعرّف على الصوتيمات اعتماداً على الشبكات العصبونية العميقة باستخدام الشبكات العصبونية التلفيفية Convolutional neural network 'CNN'. حيث نقدم طريقتين للتعرّف الطريقة الأولى المباشرة عن طريق التعرّف على الصوتيمات بمرحلة تصنيف وحيدة، وذلك بالحصول على نوع الصوتيم مباشرة عن طريق الدخل. أمّا الطريقة الثانية المقترحة تتم عن طريق عدة مراحل للتصنيف وذلك بأخذ طريقة إصدار الصوتيمات وصفوفها بعين الاعتبار (صائت vowels ونصف صائت semi-vowels وصوامت consonants و...).

واعتمدنا في الطريقتين على تحويل MelSpectrogram، حيث يجري تحويل الإشارة الصوتية إلى مصفوفة ثنائية البعد ضمن الفضاء الترددي، ومن ثم يتم إدخال هذه المصفوفة كدخل للشبكة العصبونية العميقة. قمنا باختبار المصنف المقترح على قاعدة المعطيات TIMIT، وكانت الدقّة 57% في الطريقة المباشرة، بينما حصلنا على دقة أعلى باستخدام طريقتنا المقترحة 61%.

Abstract

In the last few years, deep neural networks have taken the problem of automated voice recognition to a completely new level of accuracy. Where it provided the highest recognition rates, whether on words or on phonemes.

Voice recognition problem represents the first phase of automated speech recognition systems. In this research, we introduce the recognition of phonemes based on deep neural networks using the Convolutional neural network 'CNN'.

We will discuss two methods of recognition, the direct method by recognizing the phonemes using a single classification phase by obtaining the correct phonemes directly through the input. The second proposed method uses several phases of classification by taking into account the types of phonemes and their classes (vowels, semi-vowels, explosive, etc.). In both methods, we rely on the Mel Spectrogram transform, where the acoustic signal converted into a two-dimensional matrix within the frequency domain, this matrix, inserted as the input of the deep neural network.

We tested the proposed classifier on TIMIT database, obtained 57% accuracy in the direct method, and a higher accuracy of 61% using our proposed method.

المحتويات

النشر العلمي في الأطروحة I

الملخص III

قائمة الأشكال VII

قائمة الجداول IX

الاختصارات X

مقدمة عامة XI

أهميّة البحث XIII

الفصل الأول: معلومات نظرية 1

1.1- تمهيد 1

2.1- الأعمال المنجزة مسبقاً 1

3.1- آلية إنتاج الكلام عند الإنسان 2

4.1- معاملات ميل الترددية 5

1.4.1- استخلاص معاملات ميل الترددية 5

2.4.1- تحويل فوريه السريع 6

3.4.1- تطبيق المرشحات الترددية 7

5.1- تحويل MEL SPECTROGRAM 8

6.1- قاعدة المعطيات TIMIT 9

7.1- الخاتمة 10

الفصل الثاني: الشبكات العصبونية التلافية 11

1.2- تمهيد 11

2.2- تعريف الشبكات العصبونية التلفيفية 11

3.2- البنية الأساسية للشبكات العصبونية التلفيفية 12

1.3.2- الطبقة التلفيفية The Convolution Layer 12

5.2- خاتمة 25

الفصل الثالث: المحاكاة والنتائج العمليّة 27

1.3- تمهيد 27

2.3- تجارب أوّليّة 27

1.2.3- التجربة الأولى 27

4.3– الطريقة المباشرة **36**
1.5.3– مرحلة التصنيف الأولى 46
6.3– النتيجة **52**
7.3– معلومات تقنيّة **52**
8.3– الخاتمة **53**
9.3– الآفاق المستقبلية **54**
الملحق آ ضبط ثوابت الشبكة العصبونية في المرحلة الأولى **56**
الملحق ب ضبط ثوابت الشبكة العصبونية في المرحلة الثانية **59**
المراجع **62**
الملخص **65**

قائمة الأشكال

الشكل (1-1): بنية الجهاز الصوتي. 3
الشكل (1-2): المخطط الصندوقي لعملية استخلاص معاملات ميل الترددية. 6
الشكل (1-3): المرشحات المثلثية المستخدمة في MEL-FREQUENCY. (DO, 2015) 7
الشكل (1-4): خوارزمية حساب تحويل MEL SPECTROGRAM 9
الشكل (2-1): الرسم البياني لبعض توابع التفعيل. 15
الشكل (2-2): عدّة أشكال لمصفوفة الدخل لكل من الرمزين X,O 19
الشكل (2-3): الطبقة التلفيفية 20
الشكل (2-4): نتيجة مرور مصفوفة الدخل بالطبقة التلفيفية..... 20
الشكل (2-5): تطبيق طبقة التفعيل على خرج الطبقة التلفيفية 21
الشكل (2-6): تطبيق طبقة التجميع على خرج طبقة التفعيل 22
الشكل (2-7): نتيجة تطبيق كل من الطبقات التلفيفية والتفعيل والتجميع على صورة دخل RGB..... 22
الشكل (2-8): نتيجة المرحلة الأولى من تطبيق الشبكة العصبونيّة التلفيفية على الدخل 23
الشكل (2-9): نتيجة تطبيق مرحلتين من المراحل الأساسيّة للشبكة العصبونية التلفيفية على الدخل..... 23
الشكل (2-10): خرج طبقة التسوية..... 24
الشكل (2-11): طريقة حساب التصنيف النهائي من قبل الشبكة العصبونية ذات الاتصال الكامل 25
الشكل (3-1): رسم توضيحي لمعاملات MFCC 28
الشكل (3-2): الصور الطيفيّة SPECTROGRAM على ملف من الملفات المولّدة..... 28
الشكل (3-3): نتيجة تحويل MELSPECTROGRAM على ملف من الملفات المولّدة..... 29
الشكل (3-4): خوارزمية العمل..... 34
الشكل (3-5): نتيجة تطبيق تحويل MELSPECTROGRAM على ملف صوتي للصوتيم 'AA'..... 36
الشكل (3-6): نتيجة تطبيق تحويل MELSPECTROGRAM على ملف صوتي للصوتيم 'S' 36
الشكل (3-7): مخطط الطريقة المباشر 36
الشكل (3-8): منحني الدقّة لكل من مرحلة التدريب والاختبار..... 42
الشكل (3-9): منحني الخسارة LOSS لكل من مرحلة التدريب والاختبار..... 42
الشكل (3-10): رسم توضيحي لمصفوفة الالتباس لنتيجة التصنيف بالطريقة المباشرة 43
الشكل (3-11): خوارزمية الكشف المقترحة والمعتمدة على عدّة مراحل تصنيف 45
الشكل (3-12): رسم توضيحي لمصفوفة الالتباس الناتجة عن تمييز الصوتيمات الصوامت عن الصوتيمات الصوائت..... 47

الشكل (3-13): رسم توضيحي لمصفوفة الالتباس الناتجة عن تصنيف كل من الصوامت والصوائت إلى الصفوف الجزئيّة الموافقة 49

قائمة الجداول

الجدول (1-1): بعض الدراسات المنجزة في مجال التعرّف على الصوتيمات والنتائج التي تم الحصول عليها 2
الجدول (1-2): توزّع الصوتيمات على الصفوف الجزئيّة تبعاً لطريقة إصدار الصوتيم..... 5
الجدول (3-1): بنية الشبكة العصبونية التلفيفية الخاصّة بالتجربة الأولى 29
الجدول (3-2): نتائج التجربة الأولى على مختلف السمات 30
الجدول (3-3): نتائج التجربة الأولى على مختلف السمات 31
الجدول (3-4): بنية الشبكة العصبونية التلفيفية الخاصة بمعاملات MFCC 31
الجدول (3-5): بنية الشبكة العصبونية التلفيفية الخاصة بـ SPECTROGRAM..... 32
الجدول (3-6): بنية الشبكة العصبونية التلفيفية الخاصة بـ MELSPECTROGRAM..... 32
الجدول (3-7): نتائج التجربة الثانية 33
الجدول (3-8): نتيجة تغير عدد الطبقات الأساسية على الدقّة..... 37
الجدول (3-9): نتيجة تغير عدد الطبقات التلفيفية وطبقات التجميع على الدقّة 37
الجدول (3-10): نتيجة تغير عدد المرشّحات في كل طبقة تلافيّة على الدقّة..... 38
الجدول (3-11): نتيجة تغير متحوّلات طبقة التجميع على الدقّة 38
الجدول (3-12): نتيجة إضافة طبقة DROPOUT على الدقّة 39
الجدول (3-13): نتيجة تغير عدد الطبقات المخفيّة وعدد العصبونات في كل منها على الدقّة..... 39
الجدول (3-14): نتيجة تغير خوارزميّة الأمثلة على الدقّة..... 40
الجدول (3-15): البنية النهائيّة للشبكة العصبونيّة التلفيفية المستخدمة في الطريقة المباشرة 40
الجدول (3-16): قيم الكشف الصحيح لكل من الصوتيمات المختبرة 44
الجدول (3-17): توزّع الصوتيمات على الصفوف الجزئيّة تبعاً لطريقة إصدار الصوتيم..... 44
الجدول (3-18): بنية الشبكة العصبونيّة التلفيفية المقترحة لتصنيف الصوتيمات إلى صوامت وصوائت..... 46
الجدول (3-19): بنية الشبكة العصبونيّة التلفيفية المقترحة لتصنيف كل من الصوامت والصوائت إلى الصفوف الجزئيّة الموافقة لها..... 48
الجدول (3-20): بنية الشبكة العصبونيّة التلفيفية المقترحة لتصنيف كل من الصفوف الجزئيّة إلى الصوتيمات الموافقة 50
الجدول (3-21): نتيجة تصنيف كل من الصفوف الجزئيّة إلى الصوتيمات الموافقة لها 51
الجدول (3-22): مقارنة العتاديات المستخدمة بالنسبة للزمن اللازم للتدريب 53

الاختصارات

CNN	Convolutional Neural Network	الشبكة العصبونية التلفيفية
TIMIT	Texas Instrument Massachusetts Institute of Technology	قاعدة المعطيات الصوتية
ASR	Automatic Speech Recognition	التعرّف الآلي على الكلام
MFCC	Mel Frequency Cepstral Coefficients	معاملات ميل الطيفيّة
DNN	Deep neural Network	الشبكة العصبونيّة العميقة
HMM	Hidden Marcov Model	نموذج ماركوف المخفي
GMM	Gaussian Mixture Model	نموذج المزج الغاوصي
NN	Neural Network	شبكة عصبونية
SVM	Support Vector Machine	آلة الدعم الشعاعي
	Vowels	الصوائت
	Consonants	الصوامت
	Plosives	الأصوات الانفجارية
	Nasals	الأصوات الأنفيّة
	Fricatives	الأصوات الاحتكاكيّة
	Diphthongs	الأصوات الإدغاميّة
	Dataset	قاعدة المعطيات
DFT	Discrete Fourier Transformer	تحويل فورييه المتقطّع
DCT	Discrete Cosine Transformer	تحويل التجيب المتقطّع
DTW	Discrete Wavelet Transformer	تحويل المويجة المتقطّع
GPU	Graphical Processing Unit	وخدة معالجة الرسومات
CPU	Central Processing Unit	وحدة المعالجة المركزيّة
ReLU	Rectified Linear Unit	
	Backward Propagation	الانتشار الخلفي
FC Network	Fully Connected Network	شبكة ذات اتصال كامل
SGD	Stochastic Gradient Descent	التدرّج المنحدر العشوائي

مقدمة عامة

يعتبر الكلام أكثر طريقة طبيعيّة للتواصل بين البشر، ويسمح لهم بالتعبير عن أفكارهم ومشاعرهم. لذلك جرت دراسة الكلام البشري لسنين طويلة، مما أدّى إلى ظهور تقنيات عديدة لتسهيل التواصل بين الإنسان والآلة. تستعمل هذه التقنيات لمعالجة الأصوات البشريّة أو للتفاعل مع البشر. من أهم مجالات البحث التي يتم العمل عليها هي تركيب الكلام، وترميز الكلام، والتعرّف والتوثّق من المتكلّم، والتعرّف على الكلام آلياً Automatic Speech Recognition (ASR) [2]. يمكّن تركيب الكلام الآلة من تحويل النص المكتوب إلى كلام مفهوم. يمكن استعمال هذه التقنية من أجل الإعلانات، قراءة النص المكتوب ضمن الأجهزة وأيضاً لمساعدة من لديهم ضعف في نطق الكلام. يهدف ترميز الكلام إلى ضغط إشارة الكلام مع المحافظة على جودتها. يستعمل مثلاً في نقل الكلام عبر الانترنت Voice over Internet Protocol (VoIP) وأيضاً من أجل الاتصالات الخلوية. يستعمل تعرف المتكلم لتمييز متكلم ما من مجموعة متكلمين معرفة مسبقاً، بينما يستعمل توثق المتكلم للتأكد من صحة ادعاء شخص ما بامتلاكه هوية محددة. يستعمل التعرّف على الكلام آلياً لتحويل الكلام إلى نص مكتوب، ويكون خرج نظام تعرف الكلام هو سلسلة كلمات متوافقة مع ما قاله المتكلم. نركز في هذا البحث على تعرّف الكلام آلياً ASR، والذي يستعمل في العديد من التطبيقات العملية، التي تهدف إلى تسهيل التواصل بين الإنسان والآلة. أحد أكثر التطبيقات المشهورة هذه الأيام لأنظمة التعرف على الكلام هي المساعد الشخصي المقاد عبر الصوت voice driven personal assistant مثل "Siri" و"Cortana". تؤمّن الواجهة المقادة بالصوت طريقة تواصل طبيعية أكثر مقارنةً مع الطريقة القديمة والتي تحتاج إلى لوحة مفاتيح أو فأرة، وبالتالي تسمح للإنسان بالقيام بمكالمة رغم انشغال يديه وعينيه، كما في حالة قيادة السيارة، وبالتالي زيادة الأمان. يستخدم أيضاً نظام التعرف على الكلام آلياً في نظام الملاحة في السيارة لتحديد وجهتها.

تتألّف الإشارة الكلاميّة من وحدات صغيرة تسمّى صوتيمات Phonemes. هذه الوحدات ليس لها معنى غالباً، وإنما اجتماع هذه الوحدات ينتج وحدات على مستوى أعلى مثل الكلمات والجمل. يعتبر التعرّف على الصوتيمات جزء أساسي من التعرّف الآلي على الكلام ASR [3]. يؤدّي تطوير نظام التعرّف على الصوتيمات، وتحسين أداؤه إلى تطوير نظام التعرّف الآلي على الكلامِ ASR [1]. أخذ التعلّم العميق حيّزاً كبيراً في الآونة الأخير في مجال تعلّم الآلة، وذلك لأهميته في إيجاد حلول لمشاكل معقّدة في عدّة مجالات، سواء في الرؤية الحاسوبية ومعالجة اللغات الطبيعية... يعتبر التعلّم العميق (بمعنى استخدام شبكات عصبونيّة عميقة) الرائد في مسألة التعرّف على الصوتيمات. فالشبكات العصبونية العميقة عبارة عن شبكات عصبونية عاديّة، ولكنها تمتاز بتعدد الطبقات، فهي تجعل الشبكات العصبونية أعمق (أكثر طبقات)، وذلك عوضاً عن جعلها أوسع (أكبر في حجم الطبقة وزيادة في عدد العصبونات)، وهذا لحل المشاكل الأعقد.

ينقسم مجال التعرّف على الصوتيمات إلى أربعة مراحل: المعالجة الأوليّة، استخلاص السمات، مرحلة التصنيف، مرحلة ربط الصوتيمات المتتالية واستنتاج الكلمات. نهتم في هذا البحث بمرحلة التصنيف.

ضمن هذه الأطروحة سنتناول مايلي:

في الفصل الأوّل نستعرض الأساس النظري الذي يرتكز عليه العمل، نبدأ بالأعمال المنجزة في مجال التعرّف على الصوتيمات، وننتقل إلى معاملات ميل الترددية وتحويل MelSpectrogram وفي النهاية نأتي على ذكر قاعدة المعطيات المستخدمة في هذا العمل.

في الفصل الثاني نأتي على شرح الشبكة العصبونيّة التلفيفية، فنستعرض طبقاتها وطريقة تدريبها والثوابت والمتحوّلات الخاصّة بها. نقدّم بعدها مثال على عمل هذه الشبكة وكيفيّة تطبيقها.

في حالة المعرفة النظريّة الكافية، يفضّل الانتقال إلى الفصل الثالث مباشرةً*

يعرض الفصل الثالث في بدايته بعض التجارب المنجزة على المعطيات وعلى الشبكة العصبونية التلفيفية. بعد ذلك يتم عرض الطريقة المباشرة في التصنيف (الطريقة التقليديّة)، ونستعرض النتائج على تطبيق هذه الطريقة على قاعدة المعطيات المستخدمة. وأخيراً نعرض الطريقة المقترحة التي تقوم بتصنيف الصوتيمات تبعاً لطريقة إصدار الصوتيم، وتعتمد على عدّة مصنّفات (شبكات عصبونيّة تلافيّة)، ونستعرض النتائج الحاصلة ونقارنها مع الطريقة المباشرة.

أهميّة البحث

مع تطوّر تقنيّة ASR ازدادت التطبيقات المطلوبة في هذا المجال بشكل واضح. ولتحقيق نظام ASR متكامل لا بد من تحقيق شرطين، الأوّل هو الاستقلاليّة عن المتكلم، والثاني أن يكون النظام عام ومناسب لأي تطبيق أو مهمة. في الحقيقة من الصعب إيجاد نظام متكامل يحقق الشرطين السابقين وذلك لعدّة أسباب، منها حالة التطوّر التقني الحاليّة، وطبيعة اللغة سواء المحكية أو المستعملة في الاتصالات.

معظم الأنظمة الموجودة حاليّاً هي أنظمة تخصصيّة، تمّ تصميمها لحل مشاكل في مجال معيّن، وذلك لأنه من السهل نسبيّاً تصميم نظام بدقّة جيّدة للتعرّف على الكلام، إذا كان هدف النظام واضح وخاص. أمّا بالنسبة لنظام التعرّف على الكلام الذي نبحث عنه، لا بدّ له من أن يكون عام، وغير مقيّد بأي مجال معيّن.

إنّ أي نظام للتعرّف على الكلمات مهما اتسع، فإنّه لا يستطيع أن يغطّي كل الكلمات الموجودة في العالم (وخاصة مع عدم وجود معجم لغوي يغطّي كل الكلمات، وكل إمكانيات ورودها، ولا يعطي أيضاً قواعد ترتيبها). بينما في التعرّف على الصوتيمات، فإن المعجم الخاص بها محدود، وذلك لأنّ الصوتيمات (الوحدات الصوتية الصغيرة) الموجودة في كل لغات العالم محدودة العدد. ومن جهة أخرى، فإن الخطأ في نظام التعرّف على الكلمات سيولّد كلمة خاطئة، بينما في نظام التعرّف على الصوتيمات فسيولّد صوتيم واحد خطأ في الكلمة. لذلك فاستخدام نظام التعرّف على الصوتيمات يساعد في تقليل نسبة الخطأ في منظومة ASR. هنالك أيضاً أفضليّة أخرى عند استخدام نظام التعرّف على الصوتيمات ألا وهي السرعة، فكلما زاد فضاء الخرج (عدد الكلمات كبير جدّاً أمام عدد الصوتيمات) كلما زاد الزمن اللازم لاختيار الخرج الصحيح.

الفصل الأول

معلومات نظرية

1.1- تمهيد

نبيّن ضمن هذا الفصل الأساس النظري الذي يرتكز عليه العمل، نبدأ بذكر الأعمال المنجزة سابقاً للتعرف على الصوتيمات، ننتقل بعدها إلى لتبيين مفهوم الصوتيمات ومعاملات ميل الترددية وشرح آلية إنتاج الكلام لدى البشر، وأهمية معاملات ميل لنمذجة تصرف الأذن البشرية، وكيفية استخراجها من الكلام عبر تطبيق تقنيات معالجة الإشارة مثل تحويل فورييه وتحويل التجيب المتقطع. نذكر بعد ذلك قاعدة المعطيات التي نعتمدها في هذا البحث، وأهميتها وميزاتها.

2.1- الأعمال المنجزة مسبقاً

يعتبر التعرّف الآلي على الكلام ASR من أهم وأقدم المواضيع التي اعتنى بها الباحثون في مجال الصوت، وذلك لما لها من تطبيقات متنوعة ومهمة في مجالات عدّة نذكر منها (مساعدة الصم والبكم، المتحدث الحاسوبي الآلي، الروبوتات ذاتية التحكّم...). تمحور العمل في هذا المجال في سياقين:

- **الأوّل اعتنى بالسمات المميزة للكلام والواجب استخلاصها من الصوت:**

استُخدِمت معاملات MFCC في الورقة البحثية [4] لتطبيق مسألة التعرّف على الصوتيمات، أضيفت معاملات دلتا ودلتا دلتا لمعاملات MFCC وذلك في الورقة [5]. مع تطوّر الشبكات العصبونية بأنواعها البسيطة والعميقة، بدأ التفكير بالتعامل مع فضاء سمات أوسع، فاعتمد Zheng في [6] على الشبكة العصبونية NN لاستخلاص السمات، وبعد ذلك بدأت المحاولات على استخدام طيف الإشارة كدخل للشبكة العصبونية [7]. في هذا البحث سنعتمد على إدخال الطيف للشبكة العصبونية العميقة، ولكن بعد تقييسه بمرشحات ميل Mel Filters وهذا ما يسمى بـ MelSpectorgram.

- **الثاني اعتنى بخوارزميات التصنيف:**

في البداية، جرى استخدام نموذج HMM مع GMM، تحسب GMM احتمالات الإصدار Emission Probabilities لكل حالة من حالات نموذج ماركوف المخفي HMM [8] حيث كانت هذه الخوارزمية هي الرائدة في هذا المجال لعدة سنين. مع ظهور الشبكات العصبونية، بدأت دراسات جديدة تعتمد على معاملات MFCC كدخل للشبكة العصبونية والتي بدورها تلعب دور المصنّف MFCC معNN [9]، وجرى استخدام

معاملات MFCC مع معاملات دلتا ودلتا دلتا مع الشبكة العصبونية كمصنّف [10]. اعتمدت بعض الدراسات على مصنّف آلة الدعم الشعاعي SVM وذلك لما له من نتائج جيّدة في مسائل التصنيف [11]. ارتكز العمل في الورقة البحثية [12] على تحويل المويجة المتقطّع DTW للتمييز بين الصفوف. قدمت الشبكات العصبونية العميقة حلّاً جيّداً لمسألة التصنيف، نظراً للعدد الكبير من المتحوّلات التي تستطيع التعامل معه بكفاءة عالية. استخدم Sreenivasa في [1] نموذج HMMمع DNN وذلك بالاعتماد على السمات الموجودة مسبقا، كما تم اعتماد نموذج يرتكز على الشبكات العصبونية العميقة فقط MFCCمع DNN [13]. لاحقاً مع تطوّر الشبكات العصبونية التلفيفية وتحقيقها لنسب عالية في مجال التصنيف [1]، جرى استخدامها كمصنّف بالاعتماد على سمات MelSpectorgram مع اعتماد CNN كمصنّف، وذلك لأنّ MelSpectorgram تتضمن فضاء سمات أكبر منه في MFCC [7] وهذا ما تحتاجه الشبكات العصبونية العميقة، لتستخلص السمات الهامة من طيف الإشارة.

في الجدول (1-1) سنلخّص أهم الدراسات والنتائج على التعرّف على الصوتيمات:

الجدول (1-1): بعض الدراسات المنجزة في مجال التعرّف على الصوتيمات والنتائج التي تم الحصول عليها

النموذج المقترح	قاعدة المعطيات	الدقة
HMM+GMM+ANN [8]	TIMIT	57
DNN+HSMM [4]	TIMIT	66
HMM+GMM [1]	KANNADA	53
HMM+ANN [1]	KANNADA	57
HMM+DNN [1]	KANNADA	61
SVM [11]	TIMIT	56
CNN [9]	TIMIT	65

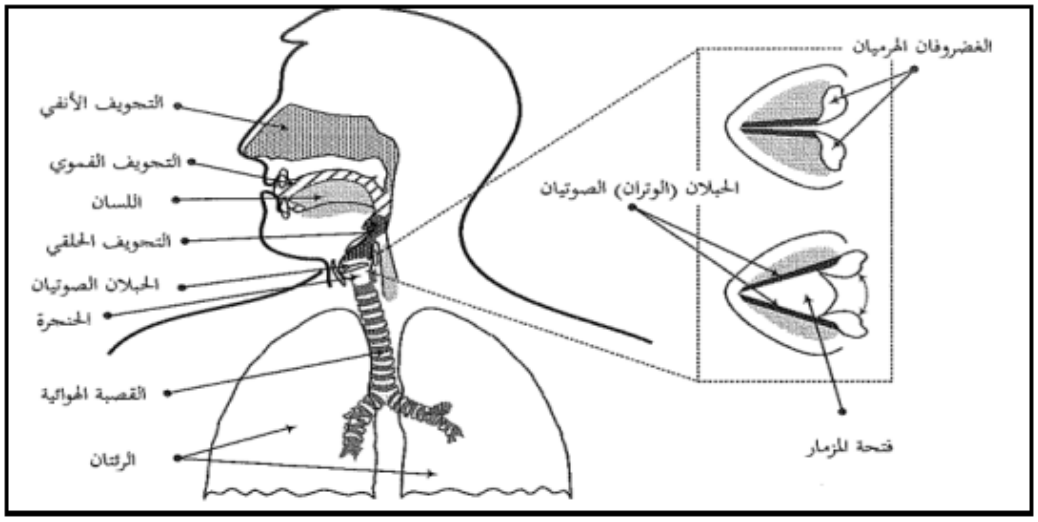
من خلال دراستنا المرجعية، نلاحظ أن نسب التعرّف على الصوتيمات –بشكل منفرد (صوتيمات مستقلّة) – على قاعدة المعطيات TIMIT تتراوح بين 45 إلى 66، عادة ما تزيد هذه النسب عند الاعتماد على قاعدة معطيات كبيرة جدّاً والاختبار على قاعدة المعطيات TIMIT [4]. وتزداد هذه النسب عند الاتجاه إلى التعرّف على سلسلة من الصوتيمات [14] حيث يوجد ثنائيات وثلاثيات من الصوتيمات والتي تتكرر بشكل واضح، كما يوجد قواعد لتسلسل الصوتيمات.

3.1- آلية إنتاج الكلام عند الإنسان

إنّ عملية النطق (إصدار الكلام) التي يقوم بها الإنسان بشكلٍ مستمرٍ هي عملية إرادية معقدة، تقوم بالأساس على تحويل الهواء الخارج من الرئتين (الزفير) إلى موجة صوتية. ولما كان نظام النطق عند الإنسان نظاماً ميكانيكياً تحكمه مجموعة من العضلات (اللسان، الحنك، الحبال الصوتية وغيرها)، ونتيجة لعطالة هذا النظام وحركته البطيئة

نسبياً يمكن اعتباره ساكناً في مدّة تتراوح بين 10 و20 ميلي ثانية، مما يسمح بنمذجته بعدّة موسطات خلال هذه المدّة، حيث تسمح النمذجة بإنتاج مختلف الأصوات.

تتدخل في عملية إنتاج الكلام معظم حواس الإنسان والجهاز التنفّسي وأحياناً العين. يعدُّ الجهاز الصوتي من أهم الأعضاء المسؤولة عن توليد الكلام عند الإنسان، ولمعرفة كيفية توليد الكلام عند الإنسان، لا بدَّ من التعرّف على أجزاء الجهاز التنفسي التي تتدخل في هذه العملية. يوضّح الشكل (1-1) بنية الجهاز الصوتي والأجزاء التي تتدخل في عملية إنتاج الكلام.



الشكل (1-1): بنية الجهاز الصوتي.

يتكوّن الجهاز الصوتي البشري من عدّة أعضاء، تشمل الحنجرة مع الوترين الصوتيين، والتجاويف الموجودة بين الحنجرة والشفتين (أي التجويف الحلقي، والتجويف الفموي، والتجويف الأنفي)، إضافةً إلى اللسان والشفتين.

ينتج الكلام عند الإنسان من اضطرابات في ضغط الهواء الخارج من الرئتين، والذي يجري بثّه عبر الجهاز الصوتي، أمّا اللاقط لهذه الإشارات فهي الأذن التي تقوم بعملية تحليل لهذه الإشارات. يجري بعد ذلك نقل النتائج إلى الدماغ الذي يقوم بتفسيرها، وبالتالي يمكن الحديث عن الكلام بأنّه فعل إرادي ومنتظم للجهاز التنفّسي والعضلي تحت إدارة النظام العصبي المركزي.

يؤمّن الجهد العضلي المبذول لإخراج الهواء من الرئتين مصدر الطاقة، حيث يخرج الهواء عبر القصبة الهوائيّة من الرئتين إلى الحنجرة، والتي تضم الحبال الصوتية. تقوم الحبال الصوتيّة بدورها بتعديل ضغط الهواء، وإذا اهتزت الحبال الصوتية، فإنها تولّد إشارات تحريض شبه دورية (تنتج الأصوات المجهورة Voiced Sounds)، حيث تكون الإشارة في المستوى الزمني شبه دورية يُسمّى دورها بالدور الأساسي Pitch Period ويُرمز له T_0، ويسمّى ترددها بالتردد الأساسي Fundamental Frequency ويُرمز بالرمز F_0. إذا لم تهتز الحبال الصوتية، فإنها تولّد إشارات ضجيجية وهكذا تنتج الأصوات المهموسة Unvoiced Sounds.

يمرُّ تيّار الهواء في التجويف الحلقي، والذي يمكن التحكّم بطوله بحسب رفع الحنجرة أو خفضها، كما يمكن تغيير مساحة مقطعه بواسطة جذر اللسان، وإمالة لسان المزمار باتجاه الجدار الحلقي الداخلي. يمرُّ بعد ذلك تيّار الهواء، إمّا عبر التجويف الأنفي أو عبر التجويف الفموي. لا يمكن التحكّم بحجم التجويف الأنفي ومواضع الأعضاء المحددة له، لذلك تقتصر وظيفته الصوتية على تعزيز نطق بعض الأصوات (مثل M,N). أمّا بالنسبة للتجويف الفموي فهو أكثر تعقيداً، حيث تتحدد أبعاده وحدوده باللسان والحنك والفكين والأسنان والشفتين، يؤدّي تغيير أبعاد وحجم هذا التجويف إلى تغيير في شكل الأمواج الصوتية بشكل بالغ التعقيد، وإلى اختلاف وتباين كبيرين في طبيعة الأصوات الصادرة أثناء الكلام. وأخيراً تأتي الشفتان، واللتان تمثلان العضو الأخير من الأعضاء التي تساهم في عمليّة النطق [15].

إنّ الرنين الصوتي–والذي يعرّف أنه تعزيز بعض الترددات في الإشارة الكلامية–الحاصل ضمن الأنبوب الصوتي، والناتج عن اهتزاز الأوتار الصوتية أثناء نطق الأصوات المجهورة، له الأثر الأساسي في تحديد خصائص الكلام. إذ يحتوي طيف هذه الإشارات على قمم يحدث عندها الرنين تسمّى البواني Formants –وهي الترددات التي يحدث عندها الرنين في الأنبوب الصوتي وتكون ذات مطال أعظمي–. أمّا في حالة الأصوات المهموسة، والتي تنتج عن تغيّر ضغط الهواء القادم من الرئتين إلى التجويف الفموي والأنفي، فتبدو هذه الأصوات في المجال الزمني مثل إشارة ضجيج وطيفها يظهر طاقة عالية عند الترددات العالية.

- **الأصوات في اللغة الانكليزيّة**

تتألّف الإشارة الكلاميّة من وحدات صغيرة تسمّى صوتيمات Phonemes. هذه الوحدات ليس لها معنى غالباً، وإنما اجتماع هذه الوحدات ينتج وحدات على مستوى أعلى مثل الكلمات والجمل. تتكوّن الإشارة الكلامية من تتالي لهذه الصوتيمات، والتي تمثّل بدورها المعلومات، أمّا ترتيب تتالي هذه الصوتيمات فهو محكوم بقواعد اللغة.

ثمّة تنوّع في الأصوات الانكليزيّة وهنالك عدّة تصنيفات لها –سنذكر التصنيف الذي سنعتمد عليه لاحقاً في خوارزميتنا المقترحة–. تقسم الصوتيمات في البداية إلى نوعين من الأصوات الصوامت Consonants والصوائت Vowels. تعرّف الصوامت بأنها الأصوات التي يتم فيها إعاقة الهواء عند خروج الصوت من الممر الصوتي، وذلك من قبل أعضاء الجهاز الصوتي، سواء بالتضييق (كما في 's')، أو بالغلق التام للشفاه (كما في 'b'). أمّا الصوائت تعرّف بأنها الأصوات التي تخرج من الجهاز الصوتي بدون أي إعاقة من أعضاء الكلام، سواء بغلق الشفاه أو تضييقها عند النطق، وهي تشبه الفتحة والضمّة في اللغة العربية، وتقسم إلى Real-Vowels, Semi-Vowels and Diphthongs.

تنتمي الصوتيمات إلى مجموعات أصغر من الصفوف الجزئية، وذلك حسب طريقة إصدار الصوت، فتقسم الصوامت إلى Plosive, Nasals and Fricatives، وبدورها تقسم الصوائت إلى Real-Vowels, Semi-Vowels and Diphthongs.

يوضّح الجدول (1-2) الصفوف الجزئيّة وطريقة إصدار كل منها [16]:

الجدول (1-2): توزّع الصوتيمات على الصفوف الجزئيّة تبعاً لطريقة إصدار الصوتيم

الصف الجزئي	الصوتيمات المنتمية	طريقة الإصدار
Plosives	b d g p t k jh ch	**إغلاق الفم لإيقاف الهواء الخارج ومن ثمّ السماح للهواء بالمرور**
Fricatives	s sh z f th v dh hh	**تضييق في عضو من أعضاء الجهاز الصوتي مما يسمح بتدفق تيار هوائي قوي في مدّة زمنية قصيرة**
Nasals	m n ng	**انفتاح التجويف الأنفي مع التجويف الفموي أثناء النطق**
Vowels	iy ih eh ae aa ah uh uw	**خروج الهواء دون وجود أي إعاقة من أعضاء الجهاز الصوتي**
Semi-Vowels	l r er w y	**شبيهة إلى حد ما بالـ Vowels ولكنّها غير مستقرة وتحدث فيها إعافة خفيفة لخروج الهواء**
Diphthongs	ey aw ay oy ow	**عن طريق دمج صائتين Tow Vowels**

4.1- معاملات ميل الترددية

معاملات ميل الترددية وتعرف أيضاً بمعاملات كيبستروم MFCC، وهي سمات تستخلص من الصوت، تهدف إلى تقليد تصرف أذن الإنسان. تعد من أهم المعاملات المستخدمة في التحليل الصوتي بسبب تقسيم المجال الترددي، وبناء عليه يتم استخدام حزم ترددية خطية في المجال الترددي أقل من 1000 Hz حيث أن الأذن البشرية حساسة لهذه الترددات المنخفضة، وحزم ترددية لوغاريتمية في المجال الترددي أعلى من 1000 Hz لأنّ الأذن ضعيفة الحساسية في هذا المجال، وبالتالي تقترب من استجابة النظام السمعي للإنسان، ويسمى هذا التقسيم في المجال الترددي بمقياس ميل الترددي Mel_frequency [17].

1.4.1- استخلاص معاملات ميل الترددية

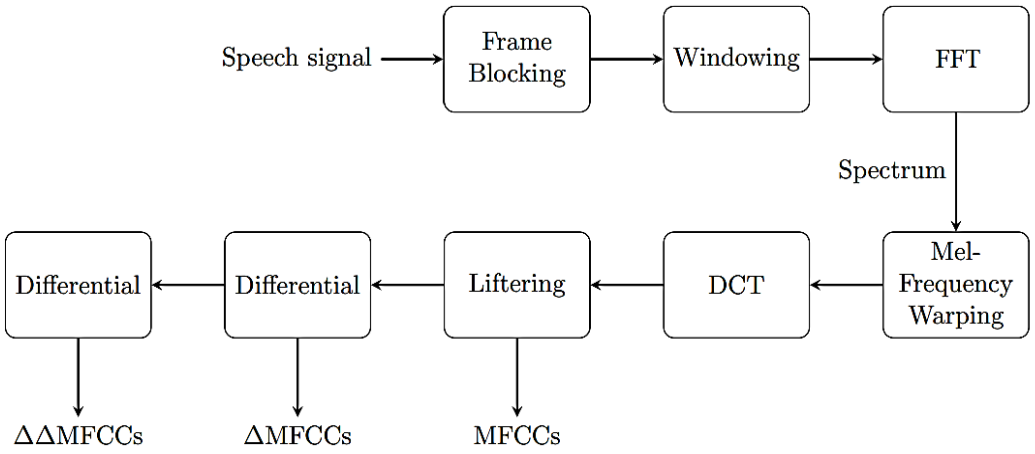
قبل أن تبدأ خوارزمية استخلاص السمات، يجب تحويل إشارة الصوت التماثلية إلى إشارة رقمية. يعاني طيف الإشارة الكلامية من تخامد بشكل ملحوظ عند الترددات العالية وذلك بسبب الإشعاع عند الشفاه. لذلك نحتاج لتعزيز هذه الترددات عبر تمرير الإشارة إلى مرشح تمرير مرتفع. تعطى علاقة هذا المرشح بالمعادلة (1-1):

$$y[n]=x[n]-k.x[n-1], \quad 0<k<1 \qquad (1-1)$$

حيث[n]y هي خرج المرشح، بينما[n]x هي الدخل. وk من رتبة 0.95 تقريباً.

تنقسم عملية استخلاص معاملات ميل الترددية MFCC إلى خمس مراحل مبينة في الشكل (1-2) وهي:

- عملية تشكيل الأطر وتقسيم الإشارة إلى إطارات.
- تشكيل النوافذ.
- عملية التحويل إلى المجال الترددي باستخدام تحويل فوريه السريع Fast Fourier Transform (FFT).
- استخدام مقياس ميل الترددي Mel_frequency.
- تحويل التجيب المتقطّع والحصول على معاملات ميل الترددية.



الشكل (1-2): المخطط الصندوقي لعملية استخلاص معاملات ميل الترددية.

سنقوم بتفصيل كل منها فيما يلي:

2.4.1- تحويل فوريه السريع

تتم عمليات تقسيم الإشارة إلى إطارات للتخفيف من أثر التداخل بتطبيق نافذة Hamming في المجال الزمني. لذلك لابد من تحويلها إلى المجال الترددي لاستخلاص المركبات الترددية المكونة لها حيث نطبق تحويل فوريه السريع –وهو خوارزمية لتطبيق تحويل فوريه المتقطع Discrete Fourier Transform (DFT) على خرج النوافذ، وحساب الطيف الترددي لكل إطار، نبقي النصف الأول من الإشارة الناتجة (الموافق للترددات الموجبة من الإشارة لأنها ستكون متناظرة كون إشارة الدخل حقيقية)، وبالتالي تهيئته ليكون دخل المرحلة التالية التي هي مرحلة المرشحات الترددية [18]. وتبين العلاقة (1-2) المعادلة الأساسية المستخدمة في حساب تحويل فوريه على N نقطة.

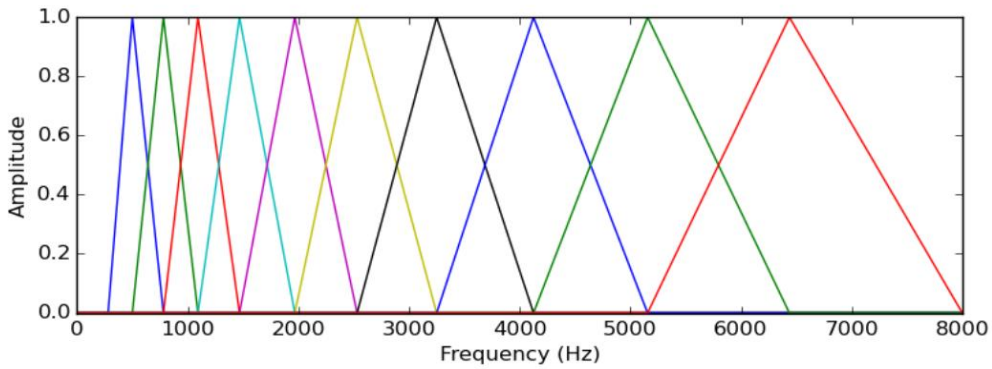
$$X_n = \sum_{k=0}^{N-1} x_k e^{\frac{-2\pi jkn}{N}}, \qquad n = 0,1,\dots,N-1 \tag{1-1}$$

حيث X_n : ناتج تحويل فوريه أي العينات الترددية للإشارة.

x_k : العينات الزمنية للإشارة.

3.4.1- تطبيق المرشحات الترددية

تستخدم المراحل الثلاث السابقة ذاتها المستخدمة في أغلب تقنيات استخراج سمات المقطع الصوتي، أما مرحلة تطبيق المرشحات الترددية Mel-Frequency Warping فتعتبر بمثابة البداية للحصول على معاملات ميل MFCC، وكما ذكرنا سابقاً إنّ هذه التقنية تقوم بمحاكاة آلية استجابة الأذن للترددات، لذلك يتم تطبيق مجموعة من المرشحات الترددية Filter Bank على كل إطار من أجل كامل الطيف الترددي لحساب استجابة الإشارة لهذه المرشحات، حيث أن الاستجابة هي حاصل ضرب عينات الإشارة الترددية بالقيم الممثلة للمرشح في المجال الترددي [19]. يمكن تمثيل هذه المرشحات بتابع رياضي ما، ولكن الأفضل أن يتم تمثيلها بتابع مثلثي كما هو موضح في الشكل (1-3).



الشكل (1-3): المرشحات المثلثية المستخدمة في **Mel-Frequency** **[20]**.

يبين الشكل (1-3) مجموعة المرشحات الترددية الخطية من أجل الحزمة الترددية أقل من 1000 Hz حيث يختلف التردد المركزي بين المرشح والآخر بمقدار ثابت. أما المرشحات في المجال الترددي أكبر من 1000 Hz فيتم تحويل التردد المركزي لكل مرشح إلى مقياس Mel بالعلاقة (1-3):

$$mel(f) = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \tag{3-1}$$

حيث يعبر f عن التردد الحقيقي و mel(f) عن التردد المتوقع في هذا المقياس.[19]

في هذا البحث تم اختيار عدد كافٍ من المرشحات 44 مرشح على كامل المجال الترددي، وذلك للحصول على تمييزية ترددية كافية.

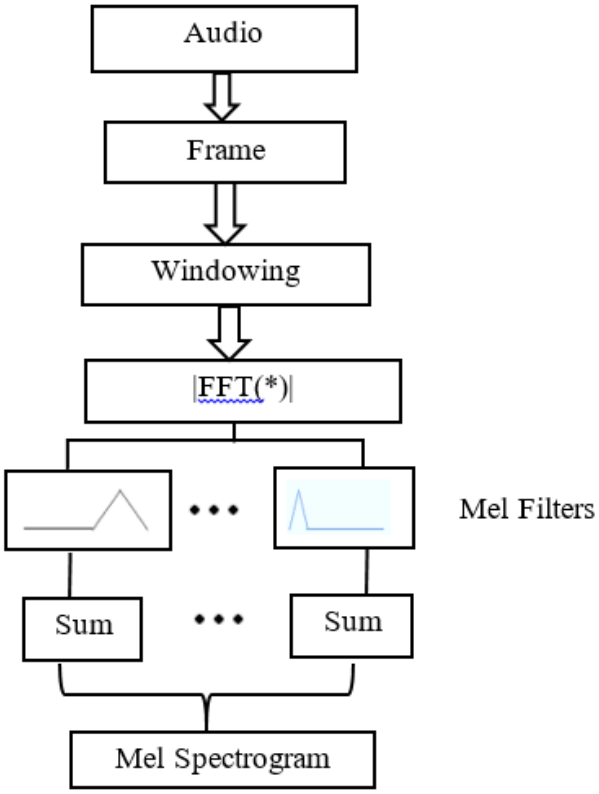
4.4.1- الحصول على معاملات كيبستروم

يقصد بكيبستروم Cepstrum تطبيق تحويل التجيب المتقطع (DCT Discrete Cosine Transform) الذي يمثل القسم الحقيقي من تحويل فورييه على خرج المرشحات الترددية لعدة أسباب، أهمها: تطبيق تحويل التجيب المتقطع DCT على الغلاف الطيفي يقوم بتركيز معظم المعلومات في الترددات المنخفضة. نستفيد من ذلك في التخلّص من تأثير شدّة الكلام وذلك بحذف المعامل الصفري للتحويل، حيث تتركز فيه الطاقة بعد تطبيق التحويل السابق، وبالتالي نتخلص من النتائج الخاطئة التي يسبّبها مطال الإشارة الصوتيّة. كما يمكن الاستفادة منه في عملية الضغط، حيث يمكننا الاكتفاء بعدد معاملات قليل. وأيضاً يجعل هذا التحويل كل معامل من معاملاته منفصلاً إحصائياً عن المعاملات الأخرى. قمنا في هذا البحث بأخذ 13 معامل فقط في المرحلة الأولى.

بعد استخراج معاملات ميل الترددية MFCC والحصول على السمات المميزة للمقطع الصوتي، لا بدّ من ذكر أن هذه المعاملات لا تحوي أيّة معلومة زمنية عن الإشارة الصوتية، أي أنّها لا تعطي أيّة معلومة عن ترابط السمات بين الإطارات المتجاورة. لذلك يمكننا الحصول على معلومات عن تغير السمات مع تقدم الإشارة زمنيّاً، فعند اشتقاق معاملات ميل الترددية MFCC نحصل على معاملات دلتا Delta وباشتقاق الأخيرة نحصل على معاملات دلتا–دلتا Delta-Delta. ما يؤدّي إلى زيادة فعالية التعرف على الإشارة الصوتيّة.

5.1- تحويل Mel Spectrogram

يتألّف اسم هذا التحويل من مفهومين، الأوّل هو الصورة الطيفية Spectrogram التي تعتبر طريقة لتمثيل الاستطاعة الطيفية لإشارة ما تبعاً لتغيراتها مع الزمن. أمّا الثاني فهو مرشحات ميل Mel (والتي سبق ذكرها) والتي هي عبارة عن مرشحات تحاكي استجابة الأذن البشرية للصوت. من خلال المفهومين السابقين يمكن صياغة التعريف التالي لتحويل Mel Spectrogram بأنّه تحويل شبيه بتحويل spectrogram المعروف ولكنّه يتميّز عنه بأنّه يحاكي استجابة الأذن عند الإنسان، وذلك عن طريق جداء نتيجة تحويل فورييه بمرشحات ميل Mel filters المحاكية لاستجابة الأذن البشرية. من جهة أخرى، يبدأ هذا التحويل كما في حساب معاملات MFCC بتقسيم الإشارة إلى إطارات زمنية، ومن ثمّ تطبيق نافذة هامينغ Hamming على كل إطار، يلي ذلك إجراء تحويل فورييه (المراحل الثلاث الأولى نفسها)[21] . نقوم بعدها بأخذ طويلة تحويل فورييه (وهنا يظهر الاختلاف بين Mel Spectrogram ومعاملات MFCC)، ومن ثمّ تطبيق مرشحات ميل على طويلة تحويل فورييه وتجميع الناتج للحصول على نتيجة تحويل Mel Spectrogram. يكمن الفرق الأساسي بين تحويل Mel Spectrogram ومعاملات MFCC بعدم وجود مرحلة تحويل التجيب المتقطّع. كما يبيّن الشكل (1–4) خوارزميّة حساب التحويل:



الشكل (1-4): خوارزمية حساب تحويل Mel Spectrogram

يعتبر هذا التحويل ذو فائدة أكبر عند التعامل مع الشبكات العصبونية العميقة [22]، فهو يحوي فضاء سمات أكبر من الموجود في معاملات MFCC، فازدياد فضاء السمات يمكّن الشبكات العصبونية العميقة من اختيار أفضل للسمات وبالنتيجة دقة أفضل. ومن جهة أخرى، يحوّل الملف الصوتي إلى فضاء ثنائي البعد (يمثل أحد المحاور الزمن والآخر يمثل التردد والشدة اللونية تمثل المطال)، فيجعل من الممكن وببساطة التعامل مع الشبكات العصبونية العميقة التي تتعامل مع الصور كدخل لها.

في عملنا، سنطبق هذا التحويل على المقاطع الصوتية الخاصة بكل صوتيم، والنتيجة ستدخل على الشبكة العصبونية العميقة (إمّا للتدريب أو للاختبار).

6.1- قاعدة المعطيات TIMIT

قاعدة معطيات صوتية تم إنشاؤها بالتعاون بين شركة Texas Instrument (TI) ومعهد Massachusetts Institute of Technology (MIT)، تحوي TIMIT على 6300 جملة صوتية مسجلة بتردد 16KHz، تعود إلى 630 متحدث (رجال ونساء وأطفال)، تحوي أيضاً على ملفات تقطيع الصوتيمات (بداية ونهاية كل صويتم، حيث تم استخراجها يدوياً). تتكوّن TIMIT من 61 صويتم ولكن يتم تجميعها واختزالها في 39 صويتم. تم اختيارنا لقاعدة المعطيات TIMIT لسببين، أوّلهما أن معظم الدراسات والبحوث الصوتية (في مجال

التعرّف على الكلام وخاصة الصوتيمات) تعتمد على TIMIT في اختبار أدائها. والثاني أنّه تم تقطيعها إلى مستوى الصوتيمات بشكل يدوي مما يعطي دقة أكبر في تحصيل الصوتيمات.

7.1- الخاتمة

قدّمنا في هذا الفصل لمحة عن بعض الأعمال المنجزة في مجال التعرّف على الصوتيمات، سواء على مستوى السمات اللازمة للتعرّف، أو على المصنّفات المستخدمة. وتعرّفنا على معاملات ميل الترددية وكيفيّة حسابها، وتعرّفنا على تحويل MelSpectrogram والذي يحاكي استجابة الأذن البشريّة للإشارة الكلاميّة. وأخيراً استعرضنا قاعدة المعطيات الصوتيّة TIMIT والتي سنعتمد عليها في هذا العمل. ننتقل في الفصل الثاني للتعرّف على الشبكات العصبونيّة التي سنستخدمها في هذا العمل.

الفصل الثاني

الشبكات العصبونية التلفيفية

1.2- تمهيد

يعرّف مفهوم تعلّم الآلة بأنّه شكل من أشكال الذكاء الصنعي، الذي يمنح الحواسب القدرة على التعلّم والتحسّن من خلال التجارب. عند تدريب خوارزميّة تعلّم الآلة عبر تزويدها بمعطيات كافية، يصبح بإمكانها تقديم التوقّعات أو حل المشكلات مثال على ذلك التعرّف على الأغراض في الصور أو الفوز في ألعاب معيّنة [23]. أمّا بالنسبة للتعلّم العميق فهو شكل محدّث من الشبكات العصبونية الصنعية، يستخدم طبقات عديدة من العصبونات لحل المشكلات الأكثر صعوبة وتعقيداً. زادت شعبيته كتقنية بشكل ملحوظ منذ عام 2005 [23]. وشجع على انتشاره فشل خوارزميات التعلّم الآلي التقليديّة في حل مهام الذكاء الصنعي المعقّدة، مثل التعرّف على الكلام أو الأغراض [24]. لقد أثبت التعلّم العميق بالفعل فائدته في مجالات مختلفة تتراوح من الرؤية الحاسوبية إلى محركات البحث، وغالباً ما ستخدم لتصنيف المعلومات كالصور أو الصوت (مثل الشبكات العصبونية التلفيفية). شهد التعلّم العميق في السنوات الأخيرة قفزة هائلة إلى الأمام [24]، وذلك نتيجة تحسين البنية الأساسيّة للبرمجيات واستخدام وحدات معالجة الرسومات Graphics Processing Units (GPUs) في تدريب الشبكات العصبونيّة.

2.2- تعريف الشبكات العصبونية التلفيفية

هي نوع خاص من الشبكات العصبونية أمامية التغذية Forward Feeding، والتي تستمد إلهامها من العمليات البيولوجية الحاصلة في الفص البصري في دماغ الكائنات الحية حيث تعتبر حل للكثير من مشاكل الرؤية الحاسوبية والذكاء الصنعي. جاء اسمها "الشبكات العصبية التلفيفية" لأنّ جداء التلاف Convolution يعدّ مرحلة أساسيّة فيها، ومنه تعرف الشبكات العصبية التلفيفية بكونها شبكات عصبية تستخدم عملية جداء التلاف عوضاً عن جداء المصفوفات العام في طبقة واحدة عل الأقل من طبقاتها.اكتسبت الشبكات العصبونية التلفيفية في السنوات الأخيرة أهمية كبيرة في مسائل التصنيف Classification، إلّا ان تاريخها يعود إلى الثمانينات [25] حيث تم تصميم نماذج لهذه الشبكات خصيصا لمعاجلة المصفوفات متعددة الأبعاد، ففي عام 1986، كان الباحثان Hubel وWiesel يفحصان القشرة البصرية للقط عندما اكتشفوا أنها تشتمل على مناطق فرعية كانت فوق

بعضها البعض لتغطية المجال البصري بأكمله، حيث تعمل هذه الطبقات كمرشحات تقوم بمعاجلة صور الإدخال والتي يتم تمريرها بعد ذلك إلى الطبقات اللاحقة[26]. في عام 1998 حاول LeCun Yann و Joshua Bagnio تصوير الخلايا العصبية في القشرة البصرية للقط، كشكل من أشكال الشبكة العصبية الاصطناعية مما أدى إلى تأسيس أول شبكة عصبونية التفافية، فكانت LENET واحدة من أولى الشبكات العصبية التلفيفية التي ساعدت في دفع مجال التعلم العميق، وقد تم تسمية هذا العمل الرائد من قبل LeCun Yann باسم LENET5 بعد العديد من المحاولات الناجحة [26]. في ذلك الوقت، تم استخدام بنية LENET بشكل أساسي لمهام التعرف على الأحرف مثل قراءة الرموز البريدية والأرقام وما إلى ذلك. في عام 2012، قام Krizhevsky [27] باستخدام CNN للفوز في تحدي ImageNet Classification حيث قام بتدريب شبكة CNN على آلاف الصور ونجح في مرحلة الاختبار بتصنيف الصور بنسبةً خطأ معقولة وفي غضون ذلك تطورت CNN بشكل ملحوظ واعتمدت لاحقا في حل العديد من مهام الرؤية الحاسوبية.

3.2- البنية الأساسية للشبكات العصبونية التلفيفية

تتألف الشبكة العصبونية التلفيفية بشكل عام من عدة طبقات مختلفة، كل منها له وظيفته الخاصة، وبحسب LENET5 يمكن تصنيف الطبقات الرئيسة لأي شبكة عصبونية التفافية في أربع مراحل: مرحلة التلاف، مرحلة التجميع، مرحلة التسوية ومرحلة الشبكة العصبونية ذات الوصل الكامل (FC (Fully Connected.

1.3.2- الطبقة التلفيفية The Convolution Layer

تعرّف عملية جداء التلاف بشكل عام بأنها عملية حسابيّة تجري على تابعين حقيقيين. لنرى الحافز من هذه العملية

الرياضية نبدأ بمثال بسيط لتابعين من الممكن أن نطبق عليهما عملية جداء التلاف.

لنفترض أننا نتعقب طائرة بحساس ليزري، يعطي حساس الليزر الخرج x(t) المعبّر عن موضع الطائرة في اللحظة t. والآن لنفترض أن قراءات الحساس تخضع بطبيعة الحال إلى ضجيج ما. لنحصل على نتائج أقل ضجيجاً وأكثر تقديراً لمكان الطائرة، نحتاج إلى أن نقوم بأخذ المتوسط لعدّة قراءات سابقة، ولكن بالطبع يجب أن يكون للقراءات الأخيرة وزن أكبر كونها ذات صلة بالخرج الحالي أكثر من القراءات القديمة. نستطيع فعل ذلك عن طريق تابع الوزن وليكن w(a) حيث a عُمر القراءة. إذا طبقنا الوسطي الموزون على جميع القراءات، نحصل على تابع جديد s يقدّم لنا تقدير تقريبي وأكثر سلاسة لمكان الطائرة يعطى بالعلاقة (2-1):

$$s(t) = \int x(a)w(t-a)da \qquad (1-2)$$

هذه العملية تدعى جداء التلاف وترمز بالنجمة: s(t)=(x*w)(t). يجب في مثالنا أن يكون w تابع توزيع احتمالي، وإلّا لن يكون الخرج موزون، ويجب على w أن يكون معدوم لكل القيم السالبة وإلّا سوف ينظر تابعنا إلى القيم المستقبلية. هذه الشروط لازمة في مثالنا ولكن في الحالة العامة لتعريف جداء التلاف فإنه صالح لأي توابع معرّفة وقابلة للتكامل.

والآن نعود إلى الهدف وهو الشبكات العصبونية التلفيفية، حيث يكون المتحوّل الأوّل (في مثالنا x) عادةً ما يمثل الدخل. بينما التابع w النواة kernel. ويسمى الخرج أحياناً بخريطة السمات Feature Map. تكون المعطيات رقمية عندما نعمل مع الأجهزة الحاسوبية، فلذلك نعتمد جداء التلاف المتقطع الموضح في المعادلة (2-2):

$$s(t) = \int x(a)w(t-a)da = \sum_{a=-\infty}^{+\infty} x(a)w(t-a) \qquad (2-2)$$

تكون مصفوفة الدخل في تطبيقات تعلّم الآلة مصفوفة متعددة الأبعاد من معطيات الدخل، بينما تكون النواة عبارة عن مصفوفة متعددة الأبعاد من المتحولات والتي يتم تعديلها وفق خوارزمية التعلّم. ولأن كل عنصر في الدخل وفي النواة يجب أن يكون محفوظ في الذاكرة بشكل منفصل، نفترض أن هذه التوابع معدومة في جميع النقاط غير المحفوظة لدينا. أي أنّه يمكننا القيام بحساب المجموع غير المنتهي السابق بحساب المجموع على مجموعة منتهية من عناصر المصفوفات. أخيراً، عادة نحتاج في مسائل الصور (مصفوفة ببعدين) إلى تعريف جداء التلاف على مصفوفات من بعدين، وبذلك نعرفه بالعلاقة (2-3):

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \qquad (3-2)$$

أو بشكل مكافئ باستخدام الخاصيّة التبادليّة:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m.n) \qquad (4-2)$$

عادة ما نستخدم العلاقة (2-4) في خوارزميات تعلّم الآلة، وذلك كون مجال المتحوّلين m وn أصغر (أبعاد النواة). الخاصية التبادلية لجداء التلاف ظهرت من كوننا قمنا بقلب النواة بالنسبة للدخل، أي بازدياد m، يزداد دليل مصفوفة الدخل، بينما بنقص دليل مصفوفة النواة. والسبب الرئيسي لقلب النواة هو للحصول على الخاصيّة التبادليّة التي تساعد الرياضيين على القيام بالبراهين الرياضية. ولكنها في خوارزميات تعلم الآلة ليست بهذه الأهمية، لذلك تطبق الكثير من مكتبات البرمجة (في مجال الشبكات العصبونية) تابع جداء الترابط Cross-Correlation (ولكن تدعوه جداء التلاف) والذي يشبه تابع جداء التلاف ولكن بدون عملية القلب للنواة، ويعرّف بالعلاقة (2-5):

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m.n) \qquad (5-2)$$

خلال هذا العمل سنلتزم بتسمية العاملين في مجال تعلّم الآلة وهو جداء التلاف.

أ. جداء التلاف الصحيح والمماثل

عندما نقوم بتطبيق نواة ذات ابعاد 3*3على مصفوفة دخل 6*6 تكون النتيجة ذات أبعاد 4*4، أي أن المصفوفة تصغر أبعادها في كل مرة نقوم بتطبيق عملية جداء التلاف عليها. هذا يعني أن عدد مرات التي يمكن لنا إجراء جداء التلاف فيها محدود، وهذا غير وارد في مسائل التعلّم العميق التي تحوي على عدد كبير من الطبقات.

بشكل عام إذا كان الدخل بأبعاد n_{in} والمرشح أو النواة ببعد f، يعطى الخرج عندها بالعلاقة:

$$n_{out} = n_{in} - f + 1 \qquad (6-2)$$

يدعى جداء التلاف المحقق للعلاقة السابقة بجداء التلاف الصحيح Valid Convolutionوهو الذي يستخدم المصفوفة الأصلية كما هي، وهناك نوع آخر من جداء التلاف يحافظ على أبعاد الدخل وهو جداء التلاف المماثل Same Convolution، ويقوم بذلك عن طريق إحاطة أو تضميد الصورة بحواف إضافية وتدعى هذه العملية إضافة الحشوة Padding. هناك عدة خيارات لقيم العناصر المستخدمة في التضميد ولكن غالباً ما نقوم بإحاطة المصفوفة بعناصر معدومة. تصبح العلاقة السابقة المعرّفة لبعد الخرج كالتالي:

$$n_{out} = n_{in} + 2 \times p - f + 1 \qquad (7-2)$$

حيث p هو عدد العناصر المضافة لكل حافة من المصفوفة، أي أنّ بُعد المصفوفة الجديد $n_{in} + 2 \times p$. ومنه لنحافظ على أبعاد المصفوف يكفي أن تتحقق العلاقة $p = \frac{f-1}{2}$ حتى يصبح $n_{out} = n_{in}$.

ب. الخطوة Stride

يقابل المرشح في التعريف السابق عند القيام بجداء التلاف جميع العناصر (عدا حواف المصفوفة) للدخل، ويكون الانتقال بين هذه العناصر بخطوة واحدة. يمكن أن نرغب بتقليل تعقيد الحساب بأن نتجاوز بعض المواضع في الدخل على حساب عدم ظهور جميع السمات الموجودة فيه. يمكننا التفكير في هذا بكونه اختزال لخرج جداء التلاف. وإذا أردنا حساب الجداء لكل s عنصر من المصفوفة، تؤول عندئذ علاقة بُعد الخرج ببُعد الدخل إلى العلاقة (2-8):

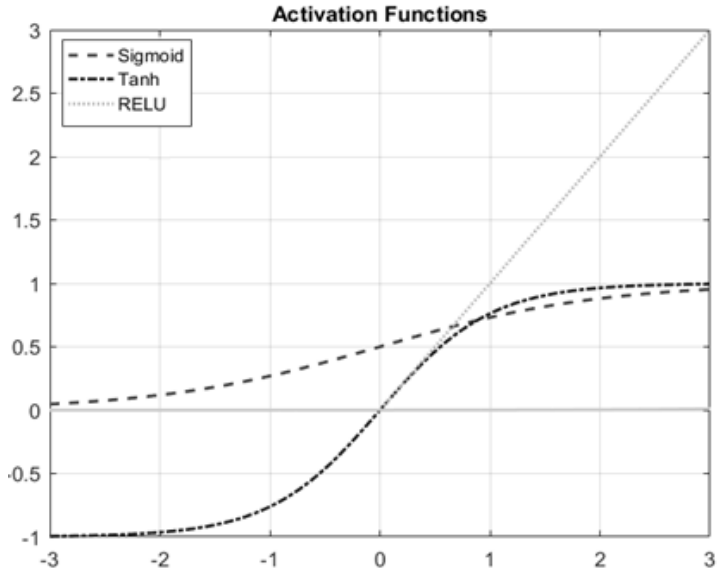
$$n_{out} = \left\lfloor \frac{n_{in} + 2 \times p - f}{s} \right\rfloor + 1 \qquad (8-2)$$

2.3.2- طبقة التفعيل

بعد انتهاء عملية التلاف يتم إدخال خريطة السمات إلى طبقة التفعيل حيث يُطبّق تابع التفعيل على كل عصبون أي ما يكافئ عنصر من خريطة السمات، ويُؤخذ بعين الاعتبار الحد من خرج العصبون وكذلك اللاخطية في عملية التفعيل حيث أنّ عمليات التلاف التي أجريت قبل هي عمليات خطية (جمع وضرب). أهم توابع التفعيل المستخدمة في هذه العملية

هو تابع RELU الذي اثبت فاعليته مقارنة بالتوابع الأخرى مثل sigmoid و Hyperbolic Tangent

الشكل (2-1) التي عانت من عدة مساوئ أهمها vanishing gradient وذلك أثناء عملية التدريب بخوارزمية الانتشار العكسي حيث يكون تابع المشتق لهذه التوابع تقريباً معدوم عند القيم الكبيرة (بالقيم المطلقة)، وهذا يبطئ عملية التدريب وحتى يمكن أن يوقفها قبل أن تنتهي.



الشكل (2-1): الرسم البياني لبعض توابع التفعيل.

لذلك يعتبر تابع ReLU الأفضل في عملية التدريب، وسنعتمد في هذا العمل على هذا التابع كتابع تفعيل. تعطى معادلة تابعReLU كما يلي:

$$F(x) = \max(0, x) \qquad (9-2)$$

حيث يتوافق خرج هذا التابع مع دخله عندما تكون قيم دخله موجبة وينعدم عند القيم السالبة.

- **مشكلة انعدام المشتق**

لا تعتبر فكرة الشبكات العصبونية العميقة فكرة جديدة من حيث المبدأ، فقد كانت مطروحة منذ زمن قديم، ولكنّها صعبة التحقيق، وذلك لأنها عانت من عدّة مشاكل. أحد هذه المشاكل التي جعلتها صعبة المنال هي

مشكلة انعدام المشتق. هذه المشكلة تجعل عمليّة التعلّم غير ممكنة أو تجعلها بطيئة جدّاً، وهذا يجعل التعامل مع الشبكات العصبونية العميقة ليس بالأمر السهل والعملي.

تظهر مشكلة انعدام المشتق نتيجة لتطبيق خوارزمية الانتشار العكسي Back-Propagation أثناء تدريب الشبكة العصبونيّة. يتم تحديث الأوزان في مختلف الطبقات مع كل تكرار Iteration، وذلك بقيمة متناسبة مع قيمة المشتق (مشتق الأوزان بالنسبة لتابع الخسارة) [29]، حيث يتم تعديل القيم وفق المشتق (2-10):

$$\frac{\partial C}{\partial W_l} \qquad (10-2)$$

حيث W_l تمثل الأوزان في الطبق ذات المرتبة l.

C تابع الخسارة في طبقة الخرج.

إنّ هذا المشتق في طبقة ما، هو الذي سيتحكّم بتحديث قيم الأوزان، وبتقريب رياضي [] يكون:

$$\frac{\partial C}{\partial W_l} \propto f'\left(Z^{(l)}\right).f'\left(Z^{(l+1)}\right).f'\left(Z^{(l+2)}\right)\ldots\ldots \qquad (11-2)$$

حيث f تابع التفعيل، $Z^{(l)}$ عصبون في الطبقة l.

تأتي مشكلة انعدام المشتق في الشبكات العصبونية، عندما تكون نتيجة $f' << 1$، وبالتالي عند جداء مجموعة من القيم أصغر بكثير من الواحد، ستصبح النتيجة قريبة إلى الصفر، وهذا يؤدي لقيمة صغيرة جدّاً للمشتق، وبالتالي ليس هنالك تحديث في قيم الأوزان.

3.3.2- طبقة التجميع Pooling

تقوم الطبقة التقليدية في الشبكة العصبية التلفيفية بثلاث خطوات لحساب الخرج لدخل ما. في الخطوة الأولى، تحسب الطبقة عدّة جداءات تلاف على التوازي لإنتاج مجموعة من القيم ندعوها بالتفعيلات الخطيّة. في الخطوة الثانية، يمر كل تفعيل خطي بتابع تفعيل غير خطي (مثل تابع ReLU) وتسمّى هذه الخطوة عادة بخطوة الكشف Detector Stage. في الخطوة الثالثة، نستخدم تابع التجميع لتعديل الخرج أكثر واختزاله.

يقوم تابع التجميع باستبدال الخرج للشبكة في موقع معين بملخّص احصائي للخرج. مثال على ذلك عملية التجميع بالأعظمي (باستخدام تابع Max Pooling) حيث يتم مقابلة كل نافذة (مجموعة من العناصر المتجاورة) بعنصر وحيد يمثل أعلى قيمة ضمن هذه النافذة. وهنالك أيضاً عمليات تجميع أخرى شهيرة مثل: عملية التجميع بالوسطي لجوار مستطيل (باستخدام تابع المتوسط) وعملية التجميع بالوسطي الموزون بناءً على البُعد عن مركز الجوار (باستخدام تابع المتوسط الموزون). يكون خرج عملية التجميع هي خريطة سمات مميزة بنفس العمق ولكن تختلف بالعرض والارتفاع.

لعملية التجميع عدة محاسن أهمها:

- تخفيض أبعاد خريطة السمات
- تخفيض عدد المتحولات والحسابات في الشبكة وهذا يساعد في التحكّم بمشكلة overfitting وهي مشكلة تحدث عندما تعطي الشبكة فعالية ممتازة في مرحلة التدريب ولكن في مرحلة الاختبار تصبح نسبة الخطأ كبيرة.
- يجعل الشبكة مقاومة لحدوث تغير او تشويه بسيط في مصفوفة الدخل (ضجيج).
- يكون مفيد عندما نهتم في معرفة إذا كانت هذه السمة موجودة أم لا عوضاً عن اهتمامنا بمكانها.

4.3.2- طبقة التسوية Flatten

بعد المرور في الطبقتين السابقتين (ولعدة مراحل) نقوم بعدها بجعل خرج المرحلتين على شكل شعاع يناسب دخل الشبكات العصبونية ليتم إدخالها إلى المرحلة الأخيرة من خوارزمية CNN.

5.3.2- طبقة الاتصال الكامل Fully Connected

تتلخص وظيفة المراحل السابقة pooling باستخلاص شعاع السمات، ففي البداية تتعلم الشبكة على اكتشاف ميزات بسيطة كالحواف مثلاً، وتستخدم هذه الحواف في الطبقة الثانية لاكتشاف الاشكال البسيطة، ثم تستخدم هذه الأشكال لكشف السمات ذات المستوى الأعلى وذلك في الطبقات الأعلى، وهكذا كلما ازداد عدد طبقات التلاف يزيد مستوى السمات التي تتعلم عليها. ليس بالضرورة أن تكون خريطة السمات مفهومة من قبل الانسان، ولكن بالنسبة للشبكة تكون شيفرة تخص صف معين (ويجري تحديد هذا الصف ضمن هذه الطبقة).

بعد استخراج السمات، يتم استخدام مصنّف لتصنيف تلك السمات وذلك باستخدام شبكات عصبونية أمامية التغذية، يكون دخلها شعاع مكوّن من خريطة السمات بعد إجراء مرحلة التجميع، وخرجها عبارة عن شعاع يعبّر عن الصف الذي تنتمي إليه خريطة السمات.

4.2- تدريب الشبكة العصبونية التلفيفية

تحوي شبكة CNN نوعين من المتحولات:

1. متحولات يتم ضبطها يدويا hyper parameter: وتتمثل بحجم وعدد الفلاتر المستخدمة في كل طبقة من طبقات التلاف بالإضافة للخطوة والحشوة وهذا كله يؤثر بشكل مباشر في حجم خريطة السمات.

2. تحولات يتم تحديثها وضبطها اثناء عملية تدريب الشبكة: وتتمثل بأوزن المرشحات أي قيم المرشحات والمتحولات المستخدمة في طبقة التصنيف وهي أوزان الشبكة العصبونية ذات الاتصال الكامل fully connected بالإضافة إلى الانحيازات.

يتم تدريب الشبكة عبر عدة خطوات:

- في البداية نقوم بإعطاء جميع الاوزان قيماً عشوائية.
- بعدها تأخذ الشبكة صورة التدريب كدخل وتمر عبر الطبقات وتحسب قيمة الخرج.
- بعد حساب قيمة الخرج نقوم بحساب تابع الخطأ والذي يعبر عن الفرق بين الخرج الصحيح وخرج الشبكة.
- بعدها نقوم باستخدام خوارزمية الانتشار العكسي التي تقوم بحساب المشتقات الجزئية لتابع الخطأ بالنسبة لجميع الأوزان، وذلك من أجل تعديل الاوزان بهدف تصغير تابع الخطأ.
- يتم إعادة الخطوة الثانية والرابعة من أجل جميع صور التدريب.

بعد التعرّف على الشبكة العصبونية التلفيفية، لابدّ أن يخطر في ذهن القارئ، لماذا نستخدم شبكة عصبونية التفافية وليس شبكة عصبونية عميقة عادية (عدّة طبقات مخفية)؟ وفيما يلي سنجيب عن هذا التساؤل:

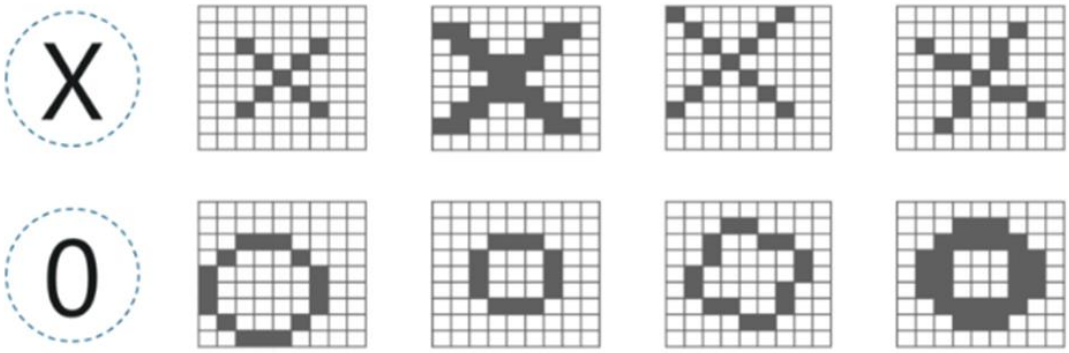
لنعتبر أنّ دخل الشبكة العصبونية العميقة العادية عبارة عن صورة $28\times28\times3$ pixels، وبالتالي سيكون في الطبقة الأولى المخفية على الأقل عدد الأوزان 2352. عادة ما تكون أبعاد صورة (أو مصفوفة) الدخل $200\times200\times3$ pixels، وهذا سيؤدّي ليكون عدد الأوزان على الأقل 120000 وهذا فقط في الطبقة المخفية الأولى، بينما سيكبر هذا الرقم بشكل كبير مع زيادة عدد الطبقات المخفية. تكمن المشكلة في زيادة عدد الأوزان في ناحيتين:

- التعقيد الحسابي مما يجعل المسألة صعبة عمليّاً
- Over-fitting زيادة التعلّم

بينما كما رأينا في الشبكة العصبونية التلفيفية، فإنها تحوي على طبقة التجميع والتي تقوم باستبدال مجموعة كبيرة من القيم بمجموعة إحصائية تعبر عنها (أقل عدداً) مثل: المتوسط أو أكبر قيمة. هذا يجعل عدد الأوزان يتناقص بشكل واضح بعد كل عملية تجميع، وبالتالي انخفاض في التعقيد الحسابي مع المحافظة على أداء عالٍ.

لفهم آلية عمل الشبكة العصبونية التلفيفية سنأخذ المثال العملي التالي والذي يهدف إلى التمييز بين صورة الرمز X وصورة الرمز O:

يوضح الشكل (2-2) صورة مثاليّة لكل من الرمزين X وO مع بعض الصور التي ممكن اعتبارها تابعة لأحد الرمزين:



الشكل (2-2): عدّة أشكال لمصفوفة الدخل لكل من الرمزين **X,O**

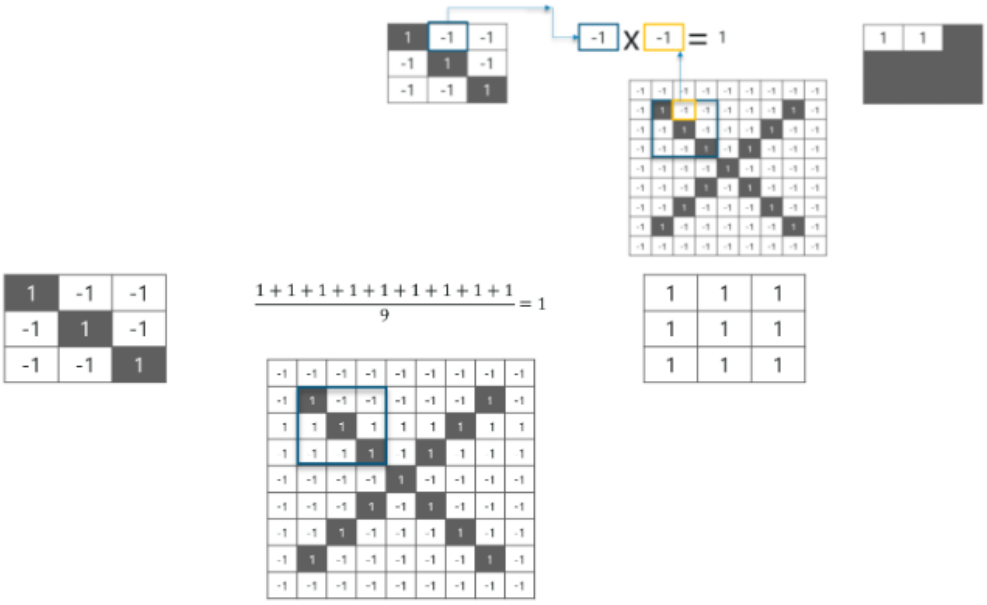
تمثل الصور السابقة بعض الصور التي يتوجب على الحاسب التعامل معها وتصنيفها، عادة تقوم البرامج الحاسوبية بالمطابقة بين صورة مرجعية وصورة الدخل (وهذا لن ينجح في هذه الحالة وذلك لأن الدخل متغيّر وغير متطابق)، في حالة طرح صورة الدخل من الصورة المرجعية ومن ثمّ أخذ عتبة مناسبة خاصة بكل رمز يمكن أن تنجح في حالة أن الرمزين مختلفين بشكل واضح (ولكن ماذا عن رموز كثيرة أو رموز متشابهة). هذا كله يدفعنا إلى عمليات أعقد واللجوء إلى الشبكات العصبونية.

أ. المرحلة الأولى (الطبقة التلفيفية)

يمثل خرج كل مرشح يتم استخدامه في هذه الطبقة مجموعة ميزات مهمة للتعرّف على الصورة (كل تطبيق للمرشح على جزء من الصورة يعتبر ميّزة)، وتقوم خوارزمية التعلّم في الشبكة العصبونية العميقة باختيار الميزات الأكثر مناسبة للصورة المرجعية.

نقوم بتمرير كل مرشح على صورة الدخل، حيث يجري في البداية جداء عناصر المصفوفة (تمثيل الصورة) الموافقة لأبعاد المرشح (في مثالنا 3×3). بعد الحصول على نتيجة الجداء، يجري قسمة الناتج على عدد عناصر المرشح (أخذ المتوسط لهذه العناصر)، عندها نكون قد حصلنا على أوّل قيمة خرج للمرشح (الميزة الأولى).

يوضح الشكل (2-3) هذه العملية:



الشكل (2-3): الطبقة التلفيفية

بتكرار العملية السابقة مع إزاحة المرشح على كامل مصفوفة الدخل، نحصل على الخرج النهائي لهذا المرشح الشكل(2-4)

0.77	-0.11	0.11	0.33	0.55	-0.11	0.33
-0.11	1.0	-0.11	0.33	-0.11	0.11	-0.11
0.11	-0.11	1.0	-0.33	0.11	-0.11	0.55
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.55	-0.11	0.11	-0.33	1.00	-0.11	0.11
-0.11	0.11	-0.11	0.33	-0.11	1.00	-0.11
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77

الشكل (2-4): نتيجة مرور مصفوفة الدخل بالطبقة التلفيفية

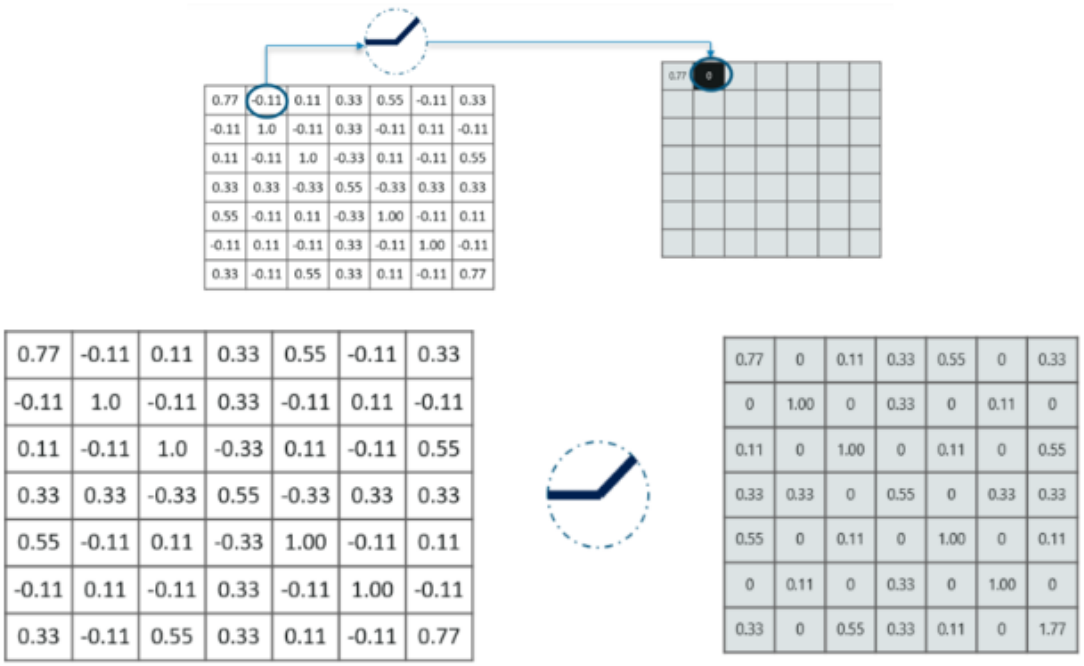
بعد الانتهاء من العمليات على المرشح الأوّل، نكرر العمليات السابقة على كل مرشحات هذه الطبقة.

ب. المرحلة الثانية طبقة التفعيل:

سنختار تابع التفعيل RELU والذي تعطى علاقته كما يلي

$$f(x) = \begin{cases} x \; if \; x \geq 0 \\ 0 \; if \; x < 0 \end{cases}$$

حيث يحافظ على القيم الموجبة كما هي، ويقوم بجعل القيم السالبة مساوية للصفر، يوضح الشكل (2-5) نتيجة تطبيق تابع التفعيل على كل عنصر من عناصر مصفوفة الدخل:



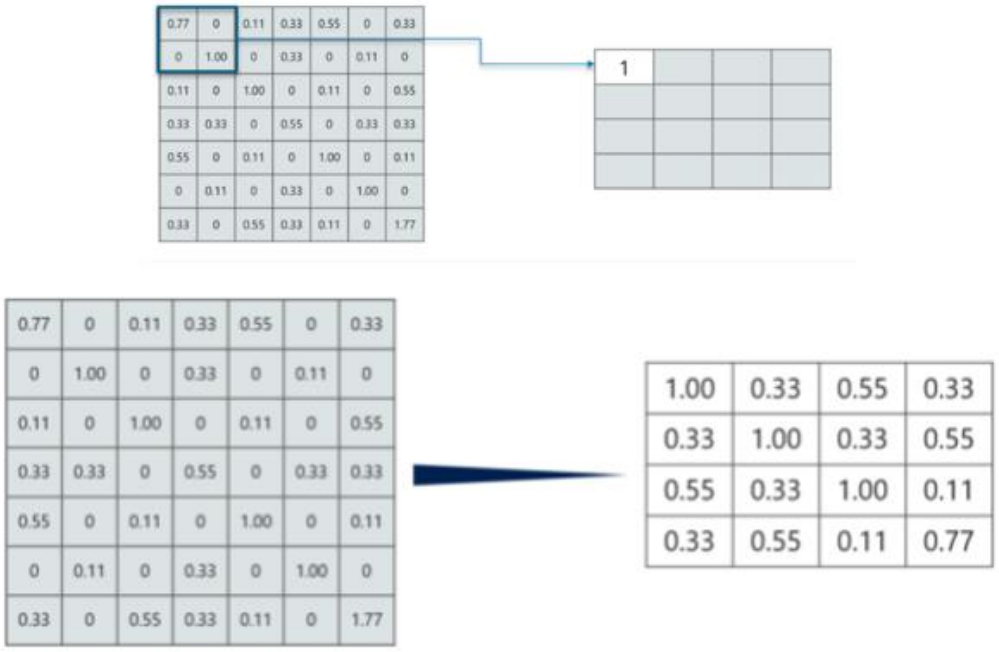
الشكل (2-5): تطبيق طبقة التفعيل على خرج الطبقة التلفيفية

ت. طبقة التجميع

تقوم باختزال خرج الطبقة السابقة، وذلك باستخدام تابع ما (مثل تابع أكبر قيمة، أو تابع المتوسط...) يعبّر عن القيم المختزلة. لتنفيذ عملية التجميع نقوم بالخطوات التالية:

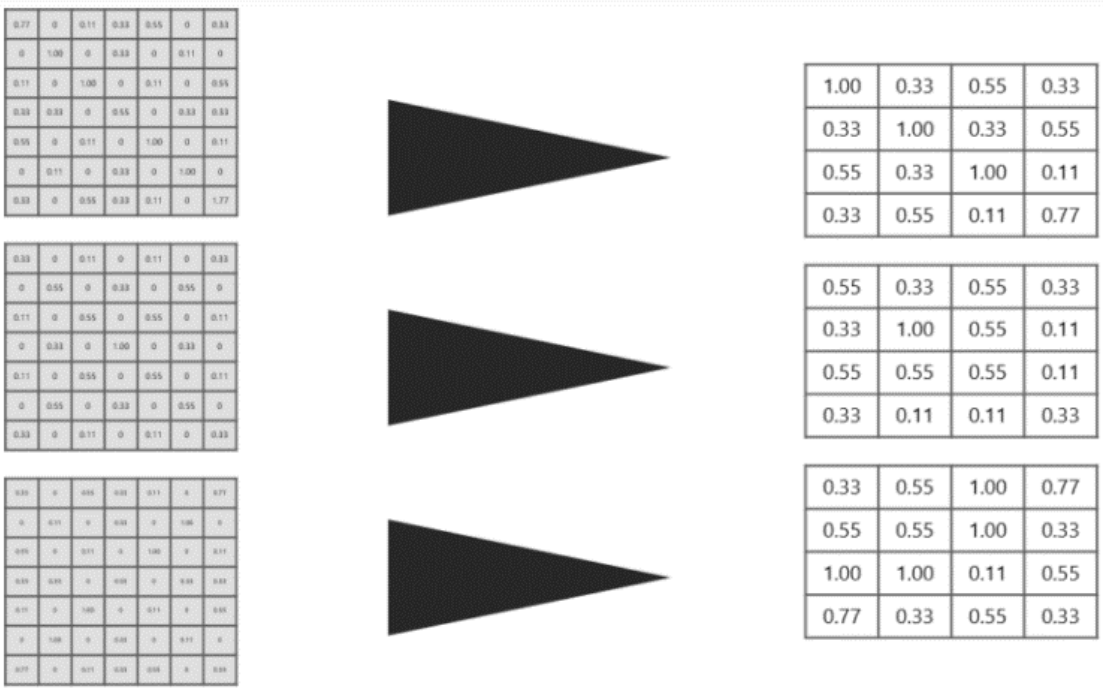
- نختار أبعاد نافذة التجميع (عادة ما يستخدم نافذة بأبعاد 2×2 أو 3×3)
- نختار طول خطوة إزاحة النافذة (في الغالب نختارها مساوية لـ 2)
- نمرر النافذة على كامل الدخل مع تطبيق تابع معيّن (في مثالنا سنستخدم تابع يعطي أكبر قيمة موجودة ضمن النافذة)

يوضح الشكل (2-6) نافذة 2×2 ونتيجة تطبيق التجميع على كامل الدخل:



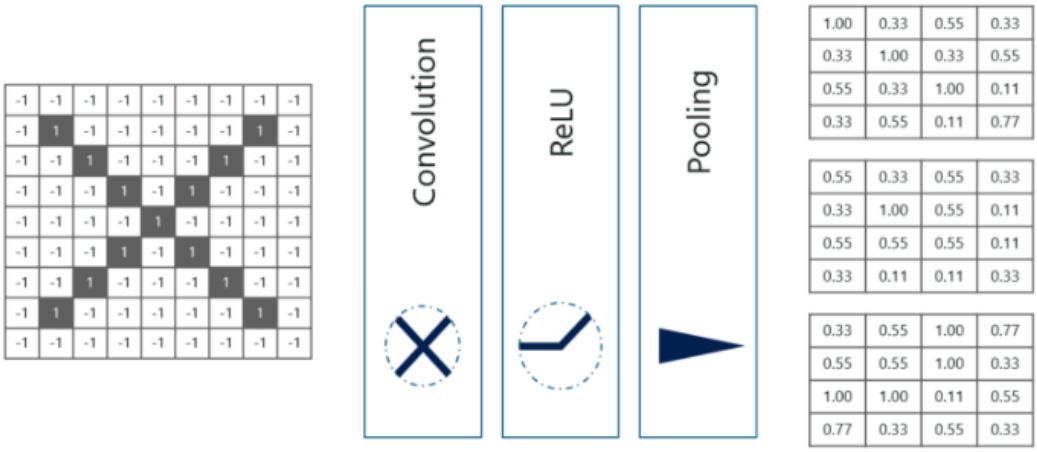
الشكل (2-6): تطبيق طبقة التجميع على خرج طبقة التفعيل

عادة ما تكون الصور ملوّنة، بمعنى لها ثلاث مركبات (ثلاث مصفوفات تعبّر عن الألوان RGB). يوضح الشكل (2-7) خرج الطبقات السابقة على المصفوفات الثلاث:

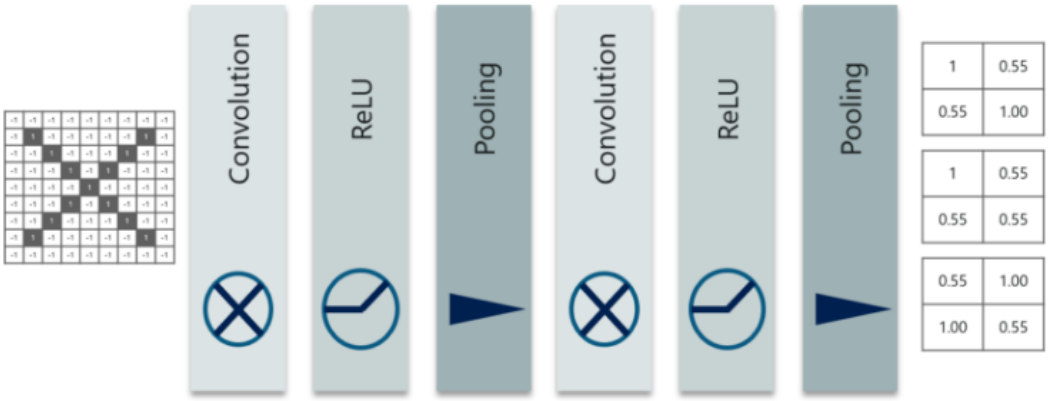


الشكل (2-7): نتيجة تطبيق كل من الطبقات التلفيفية والتفعيل والتجميع على صورة دخل **RGB**

تمثل الطبقات السابقة المرحلة الأولى في الشبكة العصبونية التلفيفية (الشكل (2-8))، وعادة ما نقوم بتكرار هذه المرحلة عدّة مرات (مع اختلاف بالمتحوّلات الخاصة بكل مرحلة). يوضح الشكل (2-9) نتيجة تطبيق مرحلتين من الطبقات على الدخل:

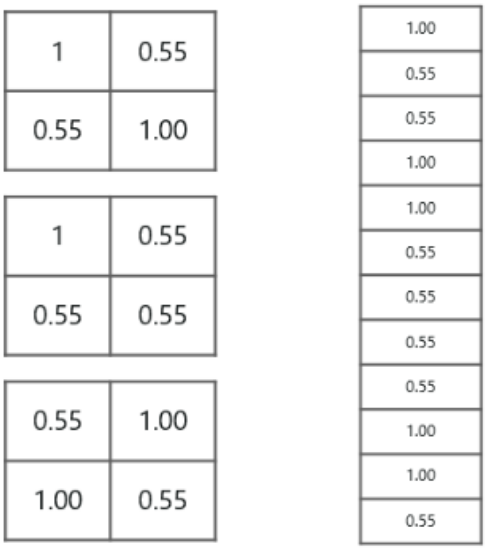


الشكل (2-8): نتيجة المرحلة الأولى من تطبيق الشبكة العصبونيّة التلفيفية على الدخل



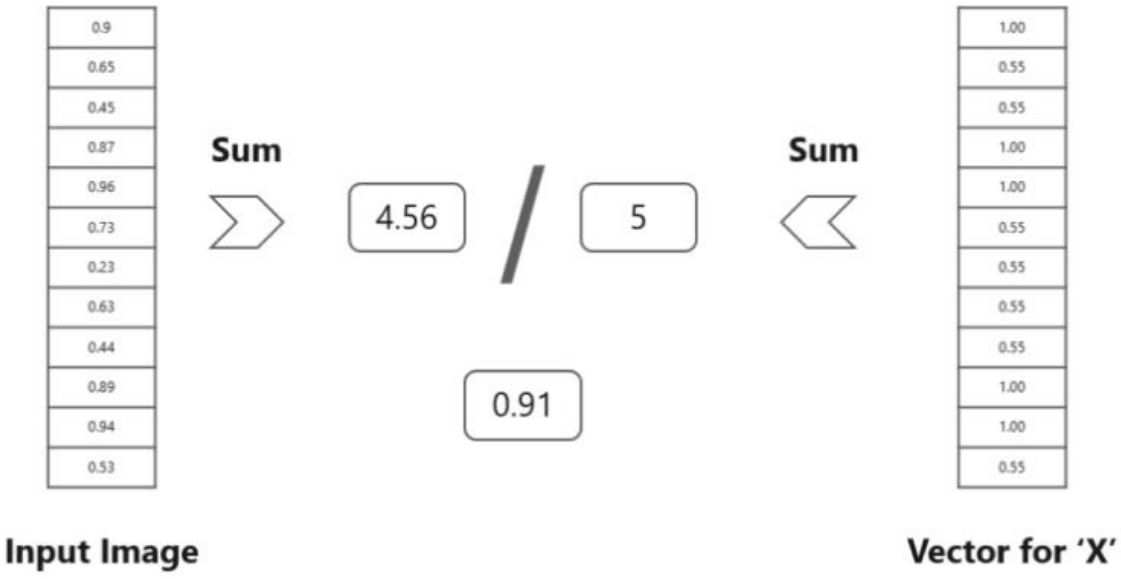
الشكل (2-9): نتيجة تطبيق مرحلتين من المراحل الأساسيّة للشبكة العصبونية التلفيفية على الدخل

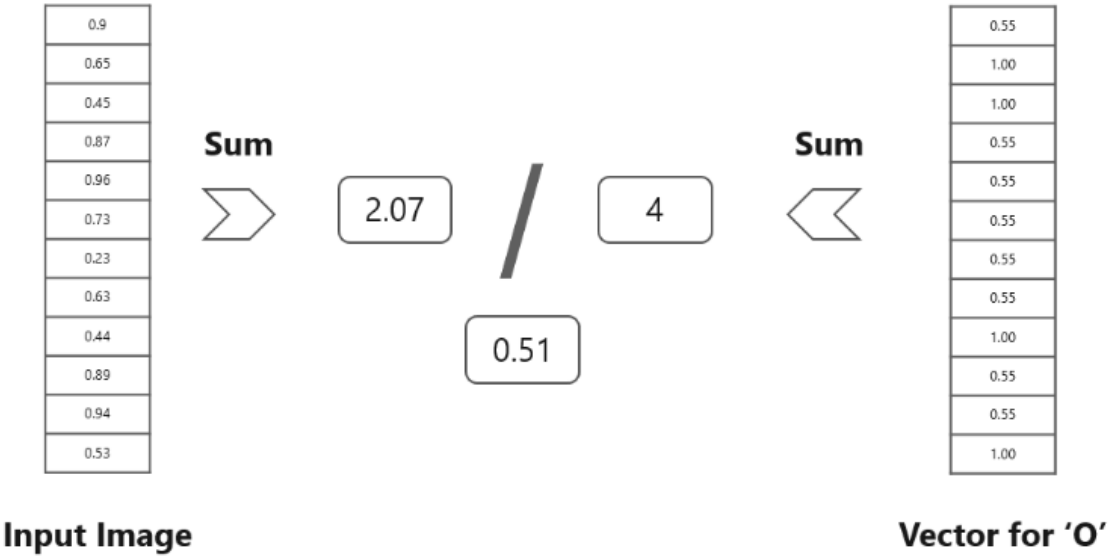
عند الانتهاء من تطبيق المراحل الأساسية، نكون قد حصلنا على ثلاث مصفوفات مختزلة تعبّر عن أهم ميزات الدخل. بعد ذلك يجري تطبيق طبقة التسوية والتي تحوّل الخرج المصفوفاتي إلى شعاع وحيد(الشكل(2-10))



الشكل (2-10): خرج طبقة التسوية

يدخل هذا الشعاع إلى **شبكة عصبونية متصلة بشكل كامل**، وذلك ليتم تهيئة الأوزان المناسبة (تدريب الشبكة)، لتقرير ما إذا كان هذا الشعاع ينتمي إلى الرمز X أو O. سنعتبر في حالتنا أن خرج الشبكة العصبونية يعبّر عن مجموع القيم (في الدخل) المقابلة لقيمة "1" في الشعاع المثالي لـ X والشعاع المثالي لـ O. عند إدخال صورة جديدة، يتم إيجاد شعاع الدخل ومن ثمّ يتم حساب قيمة مجموع العناصر (المقابلة لقيمة "1" في كل من X وO كلٍ على حدى) في هذا الشعاع ومقارنتها مع القيم المثلى لكل من الرمزين X وO، ويتم اختيار الرمز عن طريق القيمة الأقرب. يوضح الشكل (2-11) شعاع الرمز X وشعاع الدخل وشعاع الرمز O وشعاع الدخل:





الشكل (2-11): طريقة حساب التصنيف النهائي من قبل الشبكة العصبونية ذات الاتصال الكامل

من الواضح أن شعاع الدخل أقرب إلى الرمز X (0.92>0.51)، وبالتالي الصورة تعبّر عن الرمز X.

5.2- خاتمة

قدّمنا في هذا الفصل لمحة عن الشبكة العصبونية التلفيفية وأهميتها، وقمنا باستعراض الطبقات المكوّنة لها (الطبقة التلفيفية، طبقة التجميع، طبقة التفعيل، الطبقة العصبونية ذات الاتصال الكامل) ودور كل منها. وعرضنا مثال عملي يوضح آلية عمل هذه الشبكة وكيفيّة حساب خرج كل طبقة من طبقات هذه الشبكة. ننتقل في الفصل الثالث إلى استعراض الجزء العملي في هذا المشروع.

الفصل الثالث

المحاكاة والنتائج العمليّة

1.3- تمهيد

يعرض هذا الفصل في بدايته بعض التجارب المنجزة على المعطيات وعلى الشبكة العصبونية التلفيفية. تهدف هذه التجارب إلى التعرّف أكثر على الشبكة العصبونيّة التلفيفية، والتعامل معها لضبط ثوابتها بما يناسب طبيعة العمل والمعطيات. بعد ذلك يتم عرض الطريقة المباشرة في التصنيف (الطريقة التقليديّة)، والتي تعتمد على تصنيف بمرحلة وحيدة. وأخيراً نعرض الطريقة المقترحة التي تقوم بتصنيف الصوتيمات تبعاً لطريقة إصدار الصوتيم، وتعتمد على عدّة مصنّفات (شبكات عصبونيّة تلافيّة).

2.3- تجارب أوّليّة

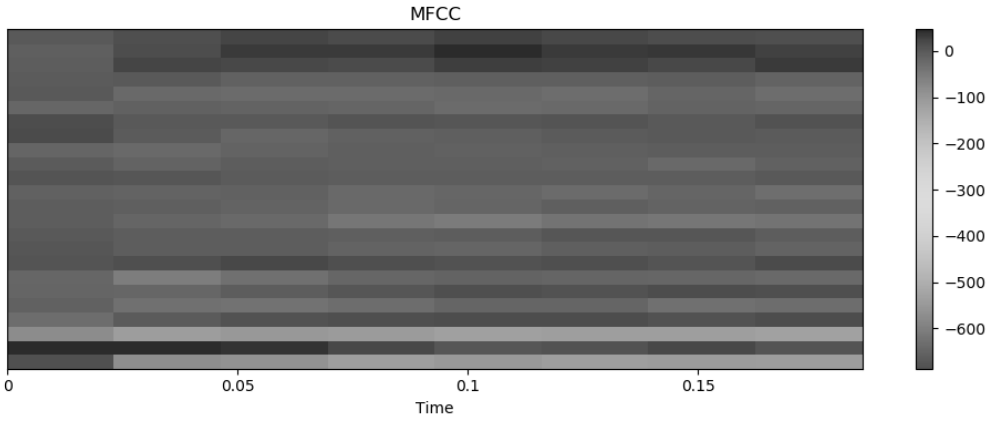
1.2.3- التجربة الأولى

ينقسم العمل في مسألة التعرّف إلى قسمين أساسيين، الأوّل يُعنى باستخلاص السمات اللازمة لعمل المصنّف بشكل جيّد، والثاني يُعنى بالمصنّف المستخدم وضبط المتحوّلات الخاصّة فيه لتحقيق أفضل تصنيف. سنقوم في هذه التجربة باختبار منظومة التعرّف وقدرتنا على إنجازها. لذلك، سنقوم بتوليد قاعدة معطيات صوتية بسيطة (إشارات جيبية)، واعتبارها دخل لمنظومة التعرّف، وذلك عن طريق تجربة ثلاثة أنواع من السمات:

- MFCC
- Spectrogram
- Mel Spectrogram

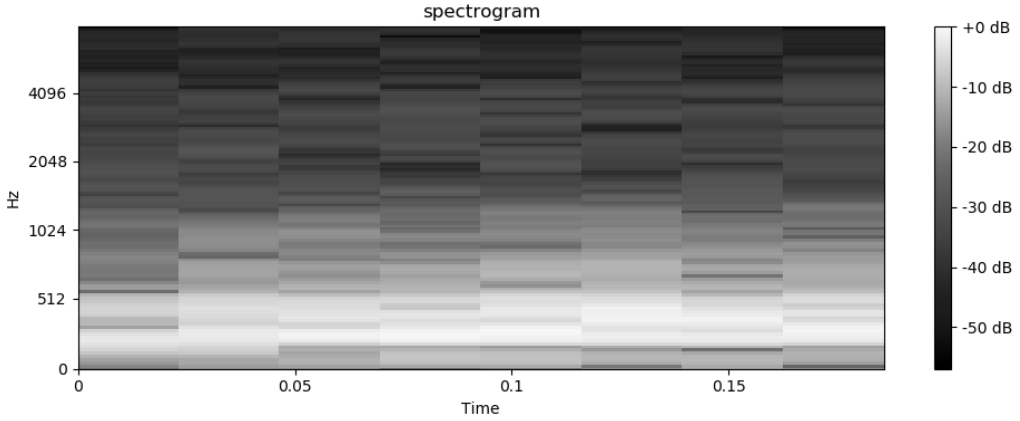
بعد استخلاص السمات سنقوم بضبط متحوّلات المصنّف (الشبكة العصبونية التلفيفية) واختباره. قاعدة المعطيات التي سنقوم بتوليدها هي عبارة عن ثماني إشارات جيبية بترددات تقع في المجال [100-240] Hz مع فرق ترددي 20Hz بين كل إشارتين متجاورتين. لكل إشارة من الإشارات الجيبية، سنقوم بتوليد 1100 ملف مختلف، وذلك عن طريق إضافة ضجيج غاوصي مختلف لكل ملف. نقوم بفرز هذه المعطيات في مجلدين (الأوّل للتدريب ويحوي 1000 ملف، والثاني للاختبار ويحوي 100 ملف). كل ملف تم تسجيله بتردد تقطيع

KHz16000 (مشابه لتردد تقطيع قاعدة المعطيات TIMIT). بعد فرز المعطيات، نقوم باستخلاص السمات من الملفات المولّدة، في البداية نقوم باستخلاص معاملات MFCC. يوضح الشكل (3-1) رسم توضيحي لمعاملات MFCC لملف وحيد من الإشارات الثمان المولّدة:



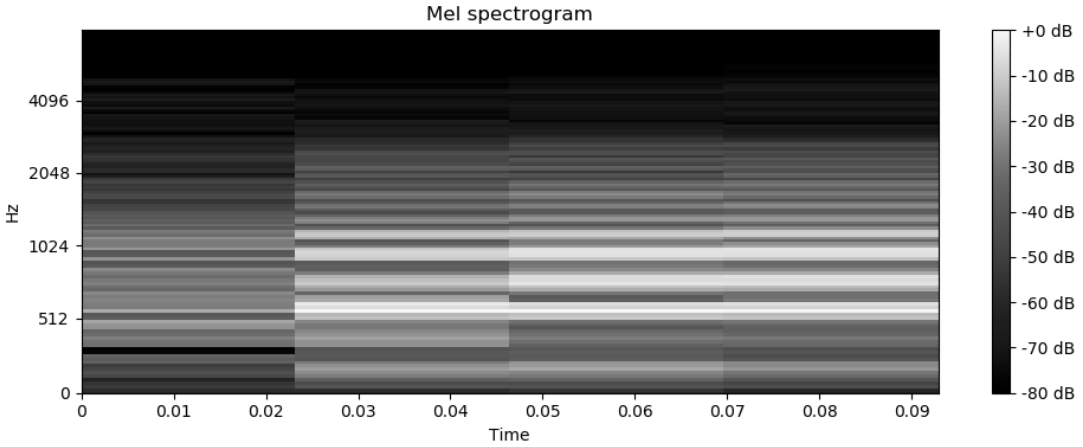
الشكل (3-1): رسم توضيحي لمعاملات **MFCC**

نقوم بعد ذلك بإجراء تحويل الصورة الطيفية Spectrogram على الملفات المولّدة. يوضح الشكل (3-2) الصورة الطيفية لبعض الإشارات المولّدة:



الشكل (3-2): الصور الطيفيّة **Spectrogram** على ملف من الملفات المولّدة

نقوم بعد ذلك بإجراء تحويل MelSpectrogram على الملفات المولّدة. يوضح الشكل (3-3) نتيجة تطبيق هذا التحويل على بعض الإشارات المولّدة:



الشكل (3-3): نتيجة تحويل **MelSpectrogram** على ملف من الملفات المولّدة

أصبح لدينا الآن ثلاثة أنواع من السمات، وبقي بناء الشبكة العصبونية التلفيفية وتدريبها على ملفات السمات واختيار أفضل السمات في مسألة التعرّف. بما أنّ الملفّات المولّدة متمايزة ترديّاً، بالتالي لسنا بحاجة إلى شبكة عصبونية عميقة جدّاً. سنستخدم للتصنيف شبكة عصبونية التفافية CNN مكوّنة من الطبقات الموضّحة في الجدول (3-1):

الجدول (3-1): بنية الشبكة العصبونية التلفيفية الخاصّة بالتجربة الأولى

نوع الطبقة	تابع التفعيل	ملاحظات
التفافية Conv1	RELU	64 مرشّح
التجميع Pool1		نواة بأبعاد 3×3
التفافية Conv2	RELU	32 مرشّح
التجميع Pool2		نواة بأبعاد 2×2
التجميع Pool3		نواة بأبعاد 2×2
التسوية Flatten		
Dropout		نسبة 20%
طبقة الدخل للشبكة العصبونية كاملة الاتصال FC Input	RELU	1024 عصبون
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden	RELU	128 عصبون
طبقة الخرج للشبكة العصبونية كاملة الاتصال FC Out	Softmax	8 عصبونات

ملاحظات على بناء الشبكة:

I. ذكرنا سابقاً أهمية اختيار تابع RELU كتابع تفعيل مع الشبكات العصبونية العميقة (ظاهرة انعدام المشتق اكتب رقم الصفحة أو الفقرة). أمّا سبب اختيار تابع التفعيل SoftMax، فتشير أدبيات الشبكات

العصبونية [30] على تفوّق تابع SoftMax على غيره من توابع التفعيل، وذلك في طبقة الخرج فقط، ومع تحقيق شرط أن يكون عدد صفوف الخرج أكبر من 2. بينما في حالة عدد الصفوف يساوي 2 فإنّ تابع التفعيل Sigmoid يعتبر الأفضل.

II. طبقة DropOut: تلعب هذه الطبقة دور كبير في منع الشبكة العصبونية من الوصول إلى حالة التعلّم الزائد OverFitting. تقوم هذه الطبقة بإطفاء عصبونات (أو جعل بعض الأوزان مساوية للصفر) بشكل عشوائي، وذلك لمنع الشبكة من الاعتماد على سمات معطيات التدريب، وذلك لكي تستطيع التعامل مع معطيات اختبار مختلفة بشكل كبير عن معطيات التدريب.

- **نتائج التجربة الأولى**

جرى اعتماد مقياس الدقة لقياس مدى صحة المصنّف (وذلك لأن أدبيات العمل في التعرّف على الصوتيمات تعتمد هذا المقياس مقارنة بمقاييس أخرى، مثل الإرجاع Recall ودقة التصنيف Precision). يتم حساب الدقة عن طريق تقسيم عدد التصنيفات الصحيحة في عينات الاختبار على عدد عينات الاختبار. يوضّح الجدول (3-2) النتائج التي حصلنا عليها في هذه التجربة:

الجدول (3-2): نتائج التجربة الأولى على مختلف السمات

السمة المستخدمة	MFCC	Spectrogram	MelSpectrogram
الدقّة %	98.7	98.4	98.9

توضح النتائج قدرة هذه الطريقة (استخدام السمات المذكورة في الجدول مع الشبكة العصبونية التلفيفية) في العمل على كشف التغيرات في ملف صوتي (الترددات)، وهذا كان الهدف الأساسي من التجربة الأولى. سنرى في التجربة الثانية، التعامل مع ملفات صوتيّة حقيقية (صوتيمات من قاعدة المعطيات)، وسنختار أفضل نوع من السمات المذكورة للتعامل معها في بقية العمل.

2.2.3- التجربة الثانية:

نعود في هذه التجربة إلى مسألتنا الأساسيّة، ألا وهي التعرّف على الصوتيمات. بداية سنختار ثماني صوتيمات من قاعدة المعطيات TIMIT، وندرّب شبكة عصبونية التفافية عليهم، وذلك باستخدام السمات الثلاثة (MFCC,Spectrogram and MelSpectrogram) لاختيار السمات الأفضل لتطبيقها على كل الصوتيمات.

تمّ اختيار الصوتيمات الثمانية بحيث تتوزع على أربعة صفوف (Vowels, Plosives, Nasals and Semi-Vowels)، وكل صوتيم سنعتمد له 2000 ملف صوتي (بين تدريب واختبار). يوضح الجدول (3-3) الصوتيمات المختارة والصف الذي تنتمي إليه:

الجدول (3-3): نتائج التجربة الأولى على مختلف السمات

الصوتيم	الصف الذي ينتمي له
aa	Vowel
ae	Vowel
b	Plosive
d	Plosive
m	Nasal
n	Nasal
r	Semi-Vowel
w	Semi-Vowel

نقوم بعد ذلك باستخلاص السمات الثلاثة، ومن ثمّ سنبني شبكة عصبونية التفافية تناسب كل نوع سمات (لتحصيل أعلى دقة لكل نوع)، وذلك لاختيار نوع السمات ذو الدقة الأعلى. يوضح الجدول (3-4) بنية الشبكة الخاصة بمعاملات MFCC، ويوضح الجدول(3-5) بنية الشبكة الخاصّة بـ Spectrogram، ويوضّح الجدول (3-6) بنية الشبكة المستخدمة مع MelSpectrogram.

الجدول (3-4): بنية الشبكة العصبونية التلفيفية الخاصة بمعاملات MFCC

نوع الطبقة	تابع التفعيل	ملاحظات
التفافية Conv1	RELU	64 مرشّح
التجميع Pool1		نواة بأبعاد 2×2 Max Pooling
Dropout		نسبة 20%
التفافية Conv2	RELU	32 مرشّح
التجميع Pool2		نواة بأبعاد 2×2 Max Pooling
Dropout		نسبة 20%
التفافية Conv3	RELU	32 مرشّح
التجميع Pool3		نواة بأبعاد 2×2 Max Pooling
التسوية Flatten		
Dropout		نسبة 20%
طبقة الدخل للشبكة العصبونية كاملة الاتصال FC Input	RELU	1024 عصبون
Dropout		نسبة 20%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden	RELU	512 عصبون
Dropout		نسبة 20%
طبقة الخرج للشبكة العصبونية كاملة الاتصال FC Out	Softmax	8 عصبونات

الجدول (3-5): بنية الشبكة العصبونية التلفيفية الخاصة بـ Spectrogram

نوع الطبقة	تابع التفعيل	ملاحظات
التفافية Conv1	RELU	64 مرشّح
التجميع Pool1		نواة بأبعاد 2×3 Mean Pooling
Dropout		نسبة 20%
التفافية Conv2	RELU	32 مرشّح
التجميع Pool2		نواة بأبعاد 2×2 Max Pooling
Dropout		نسبة 20%
التفافية Conv3	RELU	32 مرشّح
التجميع Pool3		نواة بأبعاد 2×2 Max Pooling
التسوية Flatten		
Dropout		نسبة 20%
طبقة الدخل للشبكة العصبونية كاملة الاتصال FC Input	RELU	1024 عصبون
Dropout		نسبة 20%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden	RELU	512 عصبون
Dropout		نسبة 20%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden	RELU	512 عصبون
Dropout		نسبة 20%
طبقة الخرج للشبكة العصبونية كاملة الاتصال FC Out	Softmax	8 عصبونات

الجدول (3-6): بنية الشبكة العصبونية التلفيفية الخاصة بـ MelSpectrogram

نوع الطبقة	تابع التفعيل	ملاحظات
التفافية Conv1	RELU	64 مرشّح
التجميع Pool1		نواة بأبعاد 2×3 Mean Pooling
Dropout		نسبة 20%
التفافية Conv2	RELU	32 مرشّح

التجميع Pool2		نواة بأبعاد 2×3 Mean Pooling
Dropout		نسبة 20%
التفافية Conv3	RELU	32 مرشّح
التجميع Pool3		نواة بأبعاد 2×2 Max Pooling
التسوية Flatten		
طبقة الدخل للشبكة العصبونية كاملة الاتصال FC Input	RELU	عصبونات بطول طبقة التسوية
Dropout		نسبة 20%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden	RELU	128 عصبون
Dropout		نسبة 20%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden	RELU	32 عصبون
Dropout		نسبة 20%
طبقة الخرج للشبكة العصبونية كاملة الاتصال FC Out	Softmax	8 عصبونات

نلاحظ أنّنا قمنا بزيادة عدد الطبقات في البنى الثلاث السابقة، وهذا مقارنة مع التجربة الأولى. يعود هذا الاختلاف إلى زيادة تعقيد المعطيات في التجربة الثانية.

استخدمنا في الشبكة العصبونية التلفيفية الخاصة بكل من Spectrogram وMelSpectrogram طبقة تجميع بنواة 3×2، الفائدة من استخدامها هي المحافظة على معلومات التردد، والتي تميّز الصوتيمات (حيث في كل من التحويلين يمثّل أحد المحاور الزمن والمحور الآخر التردد). كما استخدمنا تابع Mean Pooling عوضاً عن تابع Max Pooling لإجراء توسيط للقيم المتجاورة عوضاً عن أخذ أعلى قيمة. قمنا بإضافة طبقة مخفية لكل من البنى، مما يتيح قدرة أكبر على التصنيف (نتيجة لتعقيد المعطيات). البنى الثلاث المنفّذة هي أفضل بنى حصلنا عليها لكل نوع من السمات، وذلك بعد تغيير عدّة متحوّلات (كعدد المرشحات في كل طبقة، وعدد الطبقات، وعدد العصبونات ....).

- **نتائج التجربة الثانية**

بعد تدريب الشبكات الثلاث، تمّ اختبار كل شبكة على معطيات الاختبار. يوضّح الجدول (3-7) نتائج التجربة الثانية التي حصلنا عليها:

الجدول (3-7): نتائج التجربة الثانية

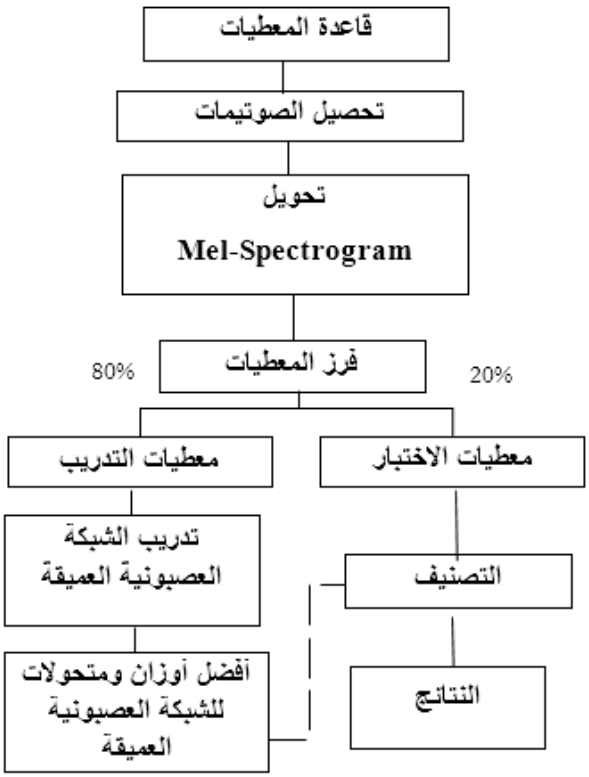
السمة المستخدمة	**MFCC**	**Spectrogram**	**MelSpectrogram**
الدقّة %	75.1	74.2	78.3

تظهر النتائج تفوّق النموذج المعتمد على تحويل MelSpectrogram على نموذج MFCC ونموذج Spectrogram. وهذا التفوّق مردّه إلى أنّ:

تحويل MelSpectrogram ذو فائدة أكبر عند التعامل مع الشبكات العصبونية العميقة، فهو يحوي فضاء سمات أكبر منه الموجود في معاملات MFCC، فازدياد حجم فضاء السمات يُمكّن الشبكات العصبونية العميقة من اختيار أفضل للسمات وبالنتيجة دقة أفضل. وبالمقابل، يؤمّن Spectrogram فضاء سمات أكبر منه في MelSpectrogram، ولكنه بحاجة إلى معطيات تدريب أكبر بكثير، حتى نتمكن من الاستفادة منه بشكل جيّد. بعد الانتهاء من التجربة الأولى والثانية، واختيار تحويل MelSpectrogram لتطبيقه على الملفات الصوتية الخاصّة بكل صوتيم. سنبدأ بالعمل مع قاعدة المعطيات TIMIT بكل صوتيماتها، وهنا سنقوم بتطبيق طريقتين:

الأولى أسميناها الطريقة المباشرة، ويتم فيها –بعد تطبيق تحويل MelSpectrogram–تصنيف الصوتيمات عبر استخدام شبكة عصبونية التفافية وحيدة. الثانية وهي طريقتنا المقترحة، ويتم فيها –بعد تطبيق تحويل MelSpectrogram–تصنيف الصوتيمات عبر عدّة مراحل تصنيف، وذلك من خلال استخدام عدّة شبكات عصبونية التفافية، تبعاً لتوزّع الصوتيمات على صفوف معيّنة.

يوضح الشكل (3-4) خوارزمية العمل العامّة لكل من الطريقتين:



الشكل (3-4): خوارزمية العمل

- يتم في البداية تحصيل الصوتيمات من قاعدة المعطيات TIMIT وفرزها (وضع كل مجموعة من المقاطع التي تتعلق بكل صوتيم ضمن مجلد منفرد).
- إجراء تحويل Mel-Spectrogram على كل مقطع صوتي وحفظ النتائج (معطيات تدريب ومعطيات اختبار).
- تدريب الشبكة العصبونية العميقة على معطيات التدريب (وحفظ الأوزان والمتحوّلات الخاصّة بالشبكة).
- اختبار الشبكة العصبونية العميقة المدرّبة على معطيات الاختبار.
- تحصيل النتائج ومقارنتها مع دراسات سابقة

3.3- التعامل مع قاعد المعطيات

ذكرنا سابقاً أنّ قاعدة المعطيات TIMIT تحوي تسجيلات صوتية لعدّة أشخاص (رجال ونساء وأطفال). بالإضافة لذلك تحوي ملفات نصيّة، توضح تقطيع الكلمات والصوتيمات (بداية ونهاية كل منها). تتيح قاعدة المعطيات هذه استخلاص الصوتيمات بشكل بسيط، وذلك عن طريق استعلام Query من قاعدة المعطيات. نقوم بتحصيل الصوتيمات اعتماداً على التابع filterdb (ضمن بيئة Matlab). توضح العلاقة التالية آلية الاستعلام:

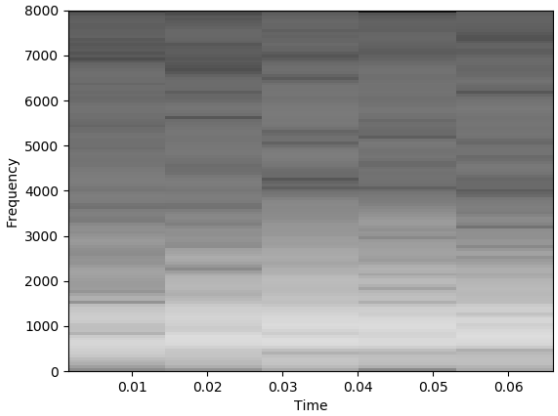
$$wavdata = filterdb(db,'phonemes', phoneme_{name}, 'ALL');$$

تدل db على قاعدة المعطيات TIMIT بعد استدعائها بتعليمه $db = ADT('Timit');$.

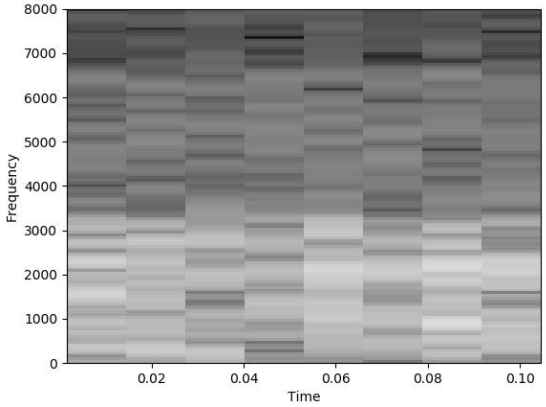
تدل $'phonemes'$ على نوع الاستعلام (هنالك استعلام أيضاً عن الكلمات 'Words').

تدل $phoneme_{name}$ على الصوتيم المراد طلب الاستعلام عنه. تدل $'ALL'$ على أنّ الطلب يتضمن كل الملفات المتعلقة بالصوتيم المطلوب.

يتم قراءة خرج الاستعلام بتابع filterdb كملف صوتي wav، ومن ثمّ يتم كتابته كملف باستخدام تعليمة audiowrite. بعد تكرار هذ العملية مع كل الصوتيمات، نكون قد خزّنا الملفات الصوتية المتعلقة بكل صوتيم في مجلّد واحد. يتم بعد ذلك فرز الملفات الصوتية المتعلّقة بكل صوتيم، ووضع كل الملفات الخاصّة بكل صوتيم بمجلّد منفرد. عند الانتهاء من فرز الصوتيمات، يتم تطبيق تحويل MelSpectrogram على كل ملف، وتخزين النتيجة في مجلّد خاص بكل صوتيم، وبهذا نكون قد هيّأنا دخل الشبكة العصبونية التلفيفية (المصنّف). يوضح الشكل (3-5) والشكل (3-6) نتيجة تطبيق تحويل MelSpectrogram على مجموعة مختلفة من الصوتيمات:



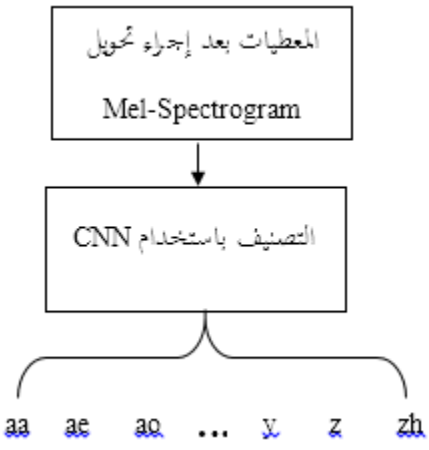
الشكل (3-5): نتيجة تطبيق تحويل **MelSpectrogram** على ملف صوتي للصوتيم **'aa'**



الشكل (3-6): نتيجة تطبيق تحويل **MelSpectrogram** على ملف صوتي للصوتيم **'s'**

4.3- الطريقة المباشرة

هي الطريقة التقليديّة في تصنيف الصوتيمات، تعتمد هذه الطريقة على تدريب شبكة عصبونية التفافية وحيدة، مهمتها تصنيف الصوتيمات بشكل مباشر. يوضح الشكل (3-7) مخطط عمل هذه الطريقة:



الشكل (3-7): مخطط الطريقة المباشرة

بدايةً، سنقوم بعدّة تجارب على متحوّلات الشبكة، وذلك لنتمكن من ضبطها بشكل جيّد.

1.4.3- ضبط ثوابت الشبكة

أ. عدد طبقات التلاف والتجميع

تمثل هذه الطبقات النواة الأساسيّة في الشبكة العصبونية التلفيفية، وتعبّر هذه الطبقات عن مرحلة استخلاص السمات من طبقة الدخل (الصوت بعد إجراء تحويل MelSpectrogram عليه). يوضّح الجدول (3-8) نتائج تغيير عدد الطبقات وتأثيره على دقّة تصنيف الصوتيمات:

الجدول (3-8): نتيجة تغير عدد الطبقات الأساسية على الدقّة

عدد الطبقات (طبقات تلاف وتجميع)	الدقّة%
2	45.3
3	45.9
4	48.2
5	45.2
6	42.2

قد يتبادر إلى الذهن أنّه كلما زدنا عدد الطبقات كلما تحسّنت الدقة، ولكن هذا التفكير غير صحيح. وذلك لأنّ زيادة عدد الطبقات تحتاج إلى زيادة في معطيات التدريب من جهة، وتحتاج إلى قدرة أكبر على ضبط العدد المتزايد من المتحوّلات (وهذا ليس بالأمر السهل أو البديهي)، بالإضافة لذلك فإن استخدام طبقات التجميع بشكل كبير قد يعمل على التأثير على السمات وحتى حذفها. لذلك سنعيد التجربة هذه ولكن مع جعل عدد طبقات التلاف أكبر من عدد طبقات التجميع. يوضّح الجدول (3-9) النتائج الجديدة لهذه التجربة:

الجدول (3-9): نتيجة تغير عدد الطبقات التلفيفية وطبقات التجميع على الدقّة

عدد الطبقات التلاف	عدد الطبقات التجميع	الدقّة%
2	1	45.2
3	1	45.9
3	2	48
4	2	48.1
4	3	50.2
5	3	53.1
5	4	51.4
6	4	46.3
6	5	44.8

نلاحظ تحسّن الدقة نسبيّاً عن حالة تساوي عدد طبقات التلاف والتجميع. كما نلاحظ أنّ حالة وجود خمس طبقات تلاف وثلاث طبقات تجميع حازت على أعلى نسبة دقّة.

ب. عدد المرشّحات في كل طبقة تلاف:

في كل طبقة تلاف نحتاج إلى تعريف عدد المرشّحات اللازمة (للقيام بعملية استخلاص السمات من مصفوفة الدخل). يعتبر عدد المرشّحات أحد ثوابت الشبكة، بينما تعتبر قيم المرشّح هي المتحوّلات التي يتم ضبطها أثناء عمليّة التدريب. يوضّح الجدول (3-10) دراسة لتغيّر عدد المرشّحات في كل طبقة مع الدقّة في تصنيف الصوتيمات:

الجدول (3-10): نتيجة تغير عدد المرشّحات في كل طبقة تلافيّة على الدقّة

عدد المرشّحات في كل طبقة	الدقّة%
[64,32,32,32,16]	49.3
[64,64,32,32,16]	50.2
[64,64,64,32,16]	48.5
[128,64,64,64,32]	52.2
[128,128,64,64,32]	54.3
[128,128,128,64,64]	50.7

نلاحظ أنّ الخيار [128,128,64,64,32] هو الخيار الأفضل في حالتنا. تمثل زيادة عدد المرشحات في الطبقة زيادة كبيرة موافقة في حجم المعطيات الواجب التعامل معها أثناء تدريب الشبكة، وهذا يؤثر على سرعة التدريب من ناحية، ومن ناحية أخرى قد تقف الموارد الحاسوبية عقبة في الزيادة الكبيرة جدّاً في عدد المرشحات.

ت. شكل طبقة التجميع:

يتوجب ضبط ثلاثة عوامل في طبقة التجميع، شكل نواة التجميع والخطوة وتابع التجميع. يوضح الجدول (3-11) دراسة لتغيير شكل النواة والخطوة وتابع التجميع مع دقّة تصنيف الصوتيمات (الطبقات الأولى في التجميع فقط، بينما الطبقات المتبقية ستكون بشكل 2×2 وذلك لتخفيف التعقيد، كما سنستخدم معها تابع MaxPooling):

الجدول (3-11): نتيجة تغير متحوّلات طبقة التجميع على الدقّة

نواة التجميع	الخطوة	تابع التجميع	الدقّة%
2×3	2	MaxPooling	49.4
2×3	2	MeanPooling	51.2
2×2	2	MaxPooling	48.2
2×2	2	MeanPooling	49.1

2×3	3	MaxPooling	48.2
2×3	3	MeanPooling	49.6
2×2	3	MaxPooling	47.2
2×2	3	MeanPooling	47.6

نلاحظ أن شكل النواة 3×2 أفضل، وهذا يناسب المعطيات التي نتعامل معها، حيث ذكرنا سابقاً أنّ معلومات تحويل MelSpectrogram موجودة في محور التردد بشكل أساسي، لذلك نحاول المحافظة على هذه المعلومات ثابتة في بداية الطبقات. بالإضافة لذلك نرى تحسّن نسبي عند استخدام تابع MeanPooling وذلك يعتبر بمثابة إجراء تنعيم على معطيات الدخل نتيجة توسيط عدّة عينات زمنية متتالية.

ث. إضافة طبقات Dropout:

ذكرنا سابقاً أن هدف هذه الطبقة هو منع حدوث حالة التعلّم الزائد over fitting، والتي تتمثل بانحياز الشبكة العصبونية لمعطيات التدريب. يوضح الجدول (3-12) نتيجة إضافة طبقة Dropout أو عدم وضعها:

الجدول (3-12): نتيجة إضافة طبقة Dropout على الدقّة

طبقة Dropout	**الدقّة%**
غير موجودة	45
عدّة طبقات (بعد طبقات التلاف والتجميع وبعد كل طبقة من طبقات FC)	49.6

ج. طبقة FC

تحوي هذه الطبقة عدة طبقات عصبونية (طبقة الدخل وطبقات مخفية وطبقة الخرج)، تحوي هذه الطبقة ثابتين يجب ضبطهما، الأوّل هو عدد الطبقات العصبونية المخفية، والثاني يمثّل عدد العصبونات في كل طبقة مخفيّة. يوضح الجدول (3-13) نتائج تغيير عدد الطبقات المخفية وعدد العصبونات في كل منها مع دقّة تصنيف الصوتيمات:

الجدول (3-13): نتيجة تغير عدد الطبقات المخفيّة وعدد العصبونات في كل منها على الدقّة

عدد الطبقات المخفية	**عدد العصبونات في كل طبقة**	**الدقّة**
1	[128]	38
1	[256]	39.7
1	[512]	40.2

2	[128 64]	44.7
2	[256 128]	44.9
2	[512 128]	44.2
3	[128 512 128]	48.5
3	[128 256 128]	51.2
3	[64 128 64]	50.4
3	[1024 512 128]	51.9
3	[1024 128 64]	54.7
4	[1024 512 128 64]	52.1
4	[512 256 256 64]	50.3

لا تعني زيادة عدد الطبقات أو العصبونات الحصول على دقّة أعلى، وخاصة إذا لم تحصل زيادة وتنوّع كبير في المعطيات، وذلك لأن النموذج سيميل بشكل واضح لملائمة معطيات التدريب (over fitting)، وبالمقابل هذا سيؤدي إلى زيادة متحوّلات الشبكة وبطء في عمليّة التعليم.

ح. تابع الأمثَلة:

قمنا باختبار عدد من خوارزميات الأمثَلة، وذلك من أجل اختيار خوارزمية الأمثَلة ذات الدقّة الأفضل. وبناءً عليه قنا باختيار خوارزمية الأمثَلة Adam بمعدّل تعلّم مساوي لـ 0.001. يوضّح الجدول (3-14) نتيجة الدقّة في التصنيف مع كل من خوارزميات الأمثَلة:

الجدول (3-14): نتيجة تغير خوارزميّة الأمثلة على الدقّة

خوارزمية الأمثَلة	**الدقّة**
Adam	56.7
Adamax	55.7
SGD	55.2

بعد الانتهاء من التجارب السابقة واختيار المتحوّلات ومعرفة الأثر التقريبي لكل منها على دقّة التصنيف، نعرض في الجدول (3-15) البنية النهائية للشبكة العصبونية التلفيفية (في حالة التصنيف المباشر):

الجدول (3-15): البنية النهائيّة للشبكة العصبونيّة التلفيفية المستخدمة في الطريقة المباشرة

نوع الطبقة	تابع التفعيل	ملاحظات
التفافية Conv1	RELU	32 مرشّح
BatchNormalization1		
التجميع Pool1		نواة بأبعاد 2×3 Mean Pooling
Dropout1		نسبة 15%

التفافية Conv2	RELU	64 مرشّح
BatchNormalization2		
التفافية Conv3	RELU	64 مرشّح
BatchNormalization3		
التجميع Pool2		Mean Pooling 2×3 نواة بأبعاد
Dropout2		نسبة 15%
التفافية Conv4	RELU	128 مرشّح
BatchNormalization4		
التفافية Conv5	RELU	128 مرشّح
BatchNormalization5		
التجميع Pool3		Mean Pooling 2×2 نواة بأبعاد
Dropout		نسبة 15%
التسوية Flatten		
طبقة الدخل للشبكة العصبونية كاملة الاتصال FC Input	RELU	عصبونات بطول طبقة التسوية
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden1	RELU	1024 عصبون
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden2	RELU	128 عصبون
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden3	RELU	64 عصبون
Dropout		نسبة 25%
طبقة الخرج للشبكة العصبونية كاملة الاتصال FC Out	Softmax	39 عصبون

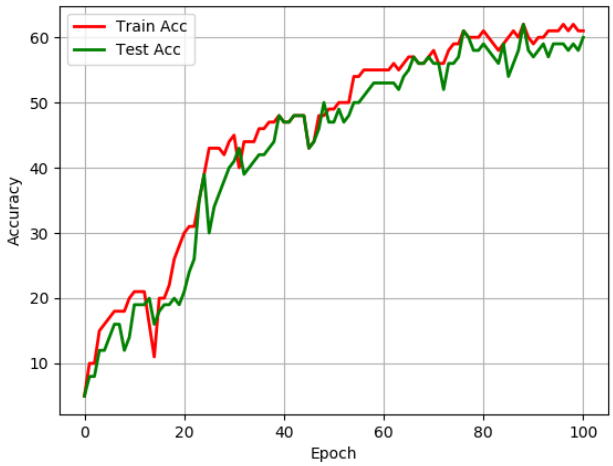
2.4.3- دور طبقة BatchNormalization

تجعل توزيع خرج توابع التفعيل لكل طبقة سابقة توزيعا منتظما. عبر تطبيق تحويل يجعل متوسط خرج توابع التفعيل قريباً من الصفر والانحراف المعياري قريباً من الواحد. تكمن الفائدة من هذه الطبقة في أمرين، الأوّل التقليل من

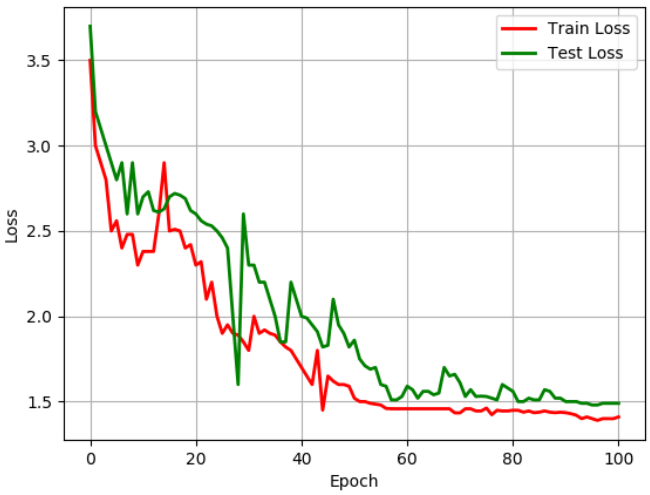
إمكانيّة حدوث التعلّم الزائد Over Fitting (عن طريق إضافة ضجيج لتوابع التفعيل)، والثاني أنها تسرّع من عمليّة التعلّم [31].

3.4.3- نتائج الطريقة المباشرة في التصنيف

بعد الانتهاء من ضبط ثوابت الشبكة العصبونية التلفيفية، نقوم الآن بتدريب الشبكة على معطيات التدريب، لذلك الأوزان ومتحوّلات الشبكة. يوضّح الشكل (3-8) والشكل (3-9) منحني الدقة ومنحني الخطأ لكل من معطيات التدريب والاختبار:



الشكل (3-8): منحني الدقّة لكل من مرحلة التدريب والاختبار



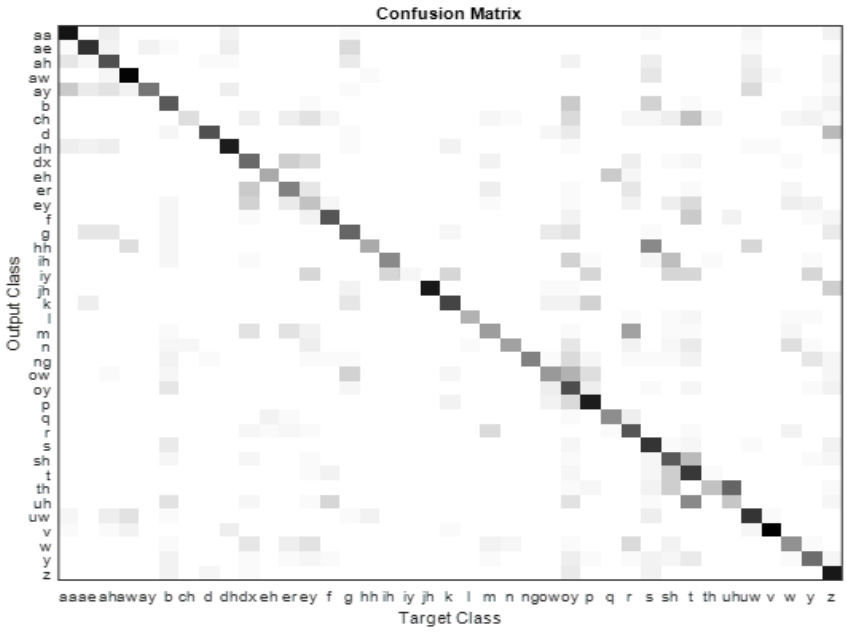
الشكل (3-9): منحني الخسارة **Loss** لكل من مرحلة التدريب والاختبار

يعطي الشكلين السابقين مؤشّر على سير عملية التدريب، حيث نلاحظ ثبات القيم تقريباً بعد التكرار رقم 80. سنقف في عملية التدريب عند التكرار رقم 100، ونحفظ الأوزان والمتحوّلات الناتجة عن التدريب.

ملاحظة:

ذكرنا في خوارزميّة العمل (الشكل (3-4)) أننا تقوم بتقسيم المعطيات إلى قسمين (التدريب والاختبار). بالنسبة لقسم التدريب، يتم التعامل معه (من قبل خوارزمية التدريب) كقسمين، قسم يتم الاعتماد عليه بالتدريب في كل تكرار، وقسم آخر يتم الاعتماد عليه في اختبار النموذج الناتج، وفي كل تكرار يتم تغيير كل من القسمين. فالشكل يدل نتائج قسم التدريب، بينما للتأكّد من صحة البنية النهائية للشبكة العصبونية التلفيفية، يتم الاعتماد على قسم الاختبار الذي لم يدخل مطلقاً إلى خوارزمية التدريب.

الآن نقوم باختبار البنية النهائيّة للشبكة العصبونية التلفيفية، وذلك بعد ضبط ثوابت ومتحوّلات هذه الشبكة. يوضّح الشكل (3-10) رسم توضيحي لمصفوفة الالتباس Confusion Matrix (تمثل هذه المصفوفة نتيجة تصنيف المعطيات الخاصّة بكل صوتيم).



الشكل (3-10): رسم توضيحي لمصفوفة الالتباس لنتيجة التصنيف بالطريقة المباشرة

تعطي مصفوفة الالتباس مؤشر على عمل الشبكة العصبونية، حيث يبيّن القطر الرئيسي فيها الكشف الصحيح للصفوف، وتمثل بقية العناصر فيها التداخلات بين الصفوف (سنوظف فيما بعد في خوارزميتنا المقترحة هذه التداخلات). لتوضيح أكثر لقيم الكشف الصحيح، نعرض الجدول (3-16) الّذي يوضّح قيم القطر الرئيسي لمصفوفة الالتباس:

الجدول (3-16): قيم الكشف الصحيح لكل من الصوتيمات المختبرة

phonemes	Aa	Ae	ah	aw	ay	b	Ch	D	Dh	Dx
accuracy	**84.7**	**75.6**	**66.5**	**90**	**54.6**	**63**	**24**	**66.3**	**82.4**	**56.9**
phonemes	**Eh**	**Er**	**ey**	**f**	**g**	**h**	**Ih**	**Iy**	**Jh**	**K**
accuracy	**39.2**	**47.1**	**31**	**64.9**	**59.4**	**38.1**	**49.2**	**5**	**82.9**	**70.4**
phonemes	**L**	**M**	**n**	**ng**	**ow**	**oy**	**P**	**Q**	**R**	**S**
accuracy	**35.8**	**41.9**	**41.6**	**41.5**	**43.6**	**67.2**	**82**	**47.9**	**64.2**	**75.3**
phonemes	**Sh**	**T**	**th**	**uh**	**uw**	**v**	**W**	**Y**	**Z**	
Accuracy	**63.5**	**73.5**	**31**	**29**	**74.4**	**91.5**	**45.7**	**57.8**	**83.4**	

ولحساب دقّة التصنيف في الشبكة المدرّبة، نقوم بجمع عناصر القطر الرئيسي في مصفوفة الالتباس، ومن ثمّ نقسمها على عدد الصفوف وهو 39 صف (صوتيم). وهكذا نكون قد حصلنا على دقّة تصنيف في هذه الخوارزميّة المباشرة 57%.

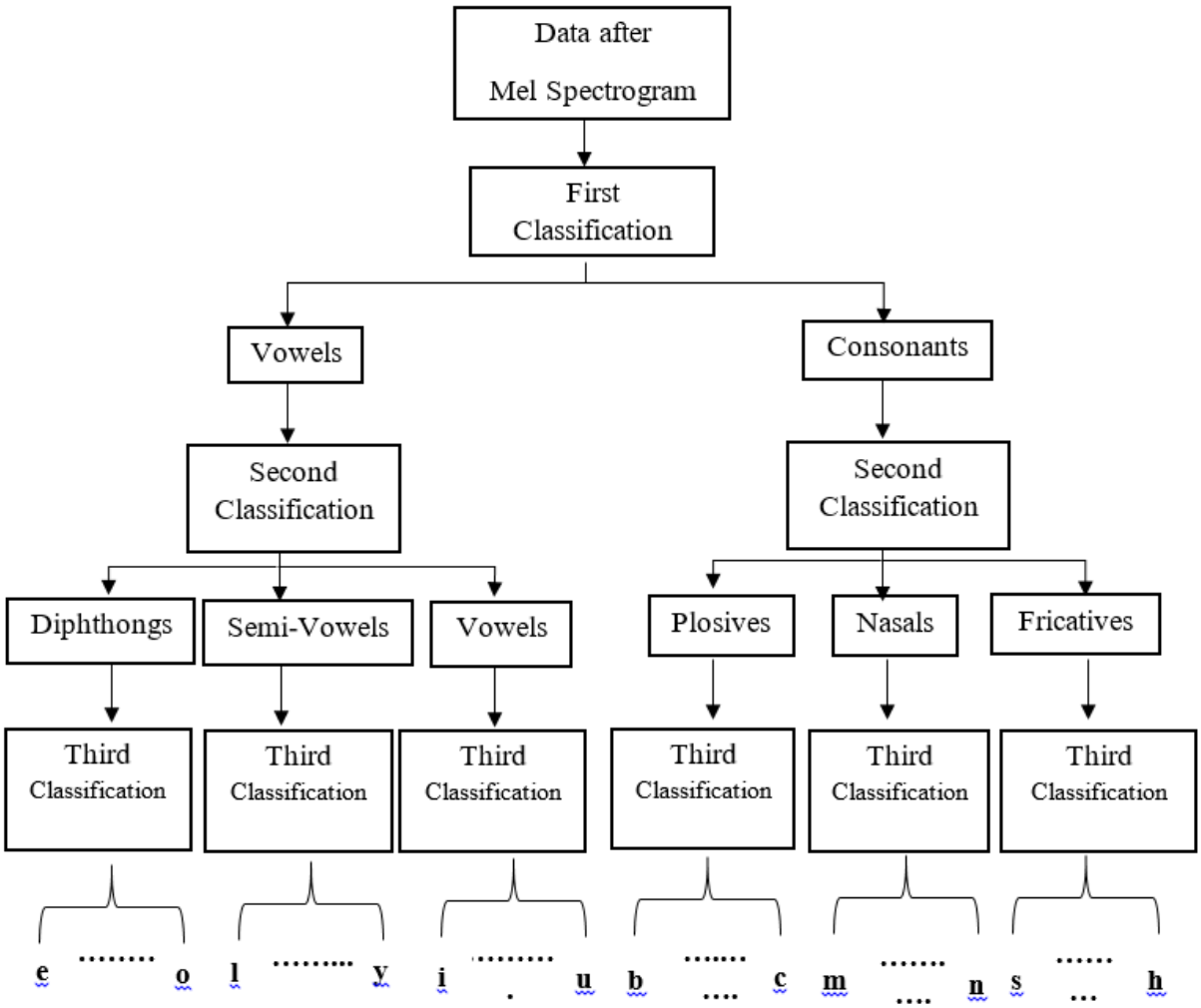
5.3- الطريقة المقترحة

كما ذكرنا سابقاً (في فقرة إنتاج الكلام)، فإنّ الصوتيمات تنتمي إلى صفوف مختلفة، وذلك حسب طريقة إنتاج الصوتيم (فم مفتوح، فم مغلق، من الحلق ....). ففي البداية تقسم الصوتيمات إلى نوعين (صفين كبيرين) وهما الصوامت Consonants والصوائت Vowels. يحوي كل من الصوامت والصوائت على مجموعة من الصفوف الجزئية، حيث ينقسم صف الصوامت إلى الصفوف الجزئية التالية (Plosives – Nasals – Fricatives)، وبدوره ينقسم صف الصوائت إلى الصفوف الجزئية التالية (Real-Vowels – Semi-Vowels – Diphthongs). سنعتمد في خوارزميتنا المقترحة على تصنيف الصوتيمات عبر عدّة مراحل باستخدام عدّة شبكات عصبونية التفافية. تعنى المرحلة الأولى بتصنيف الصوتيمات إلى صوامت وصوائت، بينما تعنى المرحلة الثانية بتصنيف كل من الصوامت والصوائت إلى الصفوف الجزئيّة الموافقة لكل منهما. أمّا المرحلة الثالثة والأخيرة، فتعنى بتصنيف كل مجموعة جزئيّة إلى الصوتيمات الموافقة لها. يوضّح الجدول (3-17) توزّع الصوتيمات على الصفوف الجزئيّة (هنالك عدّة توزّعات مختلفة للصوتيمات حسب اللهجة، ولكن سنتعامل مع هذا التوزيع لكونه معتمد في قاعدة المعطيات TIMIT [16]).

الجدول (3-17): توزّع الصوتيمات على الصفوف الجزئيّة تبعاً لطريقة إصدار الصوتيم

Secondary class	**Phonemes**
Plosives	b d g p t k jh ch
Fricatives	s sh z f th v dh hh
Nasals	m n ng
Semi-Vowels	l r er w y
Vowels	iy ih eh ae aa ah uh uw
Diphthongs	ey aw ay oy ow

نعرض في الشكل (3-11) آليّة التصنيف المتعدد للصوتيمات والمعتمدة في خوارزميتنا المقترحة، وذلك لفهم الخوارزميّة بشكل أفضل:



الشكل (3-11): خوارزمية الكشف المقترحة والمعتمدة على عدّة مراحل تصنيف

ملاحظة:

تمّ ضبط كل الشبكات العصبونية المستخدمة في الخوارزمية المقترحة (ضبط الثوابت)، وذلك بنفس طريقة ضبط الشبكة العصبونية في الطريقة المباشرة (ولمعلومات أكثر يمكن الاطّلاع على الملحق).

1.5.3- مرحلة التصنيف الأولى

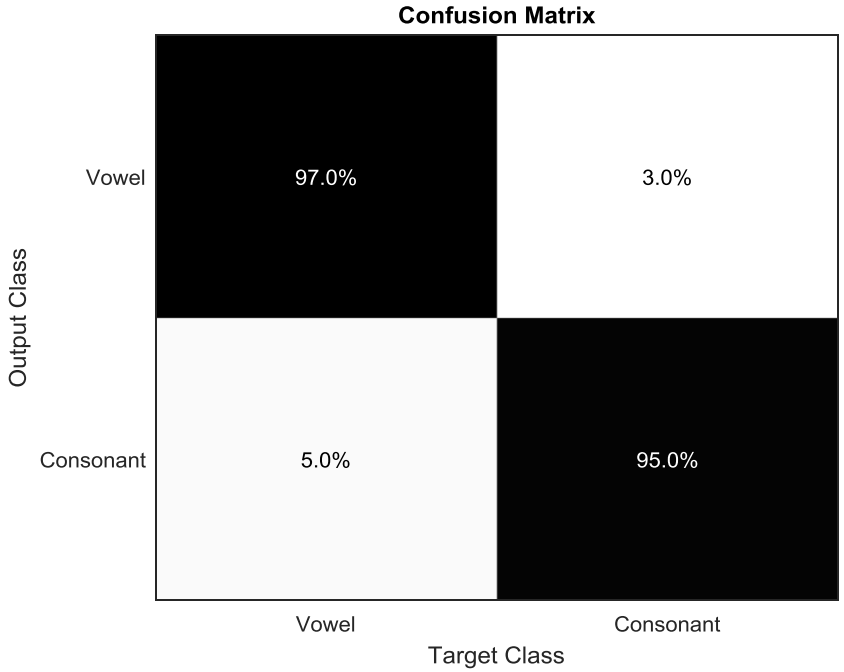
سنقوم في هذا الجزء باستخدام الشبكة العصبونيّة التلفيفية، وذلك بعد ضبط ثوابتها وتدريبها لضبط متحوّلاتها وأوزانها على تصنيف الصوتيمات إلى صوامت وصوائت. يوضّح الجدول (3-18) بنية الشبكة العصبونيّة التلفيفية المقترحة:

الجدول (3-18): بنية الشبكة العصبونيّة التلفيفية المقترحة لتصنيف الصوتيمات إلى صوامت وصوائت

نوع الطبقة	تابع التفعيل	ملاحظات
التفافية Conv1	RELU	32 مرشّح
BatchNormalization1		
التجميع Pool1		نواة بأبعاد 2×3 Mean Pooling
Dropout1		نسبة 15%
التفافية Conv2	RELU	64 مرشّح
BatchNormalization2		
التجميع Pool2		نواة بأبعاد 2×3 Mean Pooling
Dropout2		نسبة 15%
التفافية Conv3	RELU	128 مرشّح
BatchNormalization4		
التجميع Pool3		نواة بأبعاد 2×2 Max Pooling
Dropout		نسبة 15%
التسوية Flatten		
طبقة الدخل للشبكة العصبونية كاملة الاتصال FC Input	RELU	عصبونات بطول طبقة التسوية
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden1	RELU	128 عصبون
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden2	RELU	32 عصبون
Dropout		نسبة 25%
طبقة الخرج للشبكة العصبونية كاملة الاتصال FC Out	sigmoid	2 عصبون

نلاحظ أنّ هذه الشبكة أقل عمقاً، وذلك مقارنة مع الشبكة المستخدمة في الطريقة المباشرة، ويعود ذلك إلى أنّ الصوامت والصوائت أكثر اختلافاً (مقارنةً بالاختلاف بين كل الصوتيمات). كما نلاحظ أننا استخدمنا تابع التفعيل Sigmoid في طبقة الخرج، وذلك لأن عدد صفوف الخرج صفين (الصوامت والصوائت).

بعد الانتهاء من تدريب الشبكة العصبونيّة لضبط المتحوّلات والأوزان الخاصّة بالشبكة، نقوم باختبار المصنّف على معطيات الاختبار. يوضح الشكل (3-12) رسم توضيحي لمصفوفة الالتباس الخاصّة بالمرحلة الأولى من التصنيف:



الشكل (3-12): رسم توضيحي لمصفوفة الالتباس الناتجة عن تمييز الصوتيمات الصوامت عن الصوتيمات الصوائت

توضح مصفوفة الالتباس حصولنا على دقّة تصنيف 96% في المرحلة الأولى (هذه الدقّة هي المتوسط الحسابي لكل من دقتي الصوامت 95 والصوائت 97).

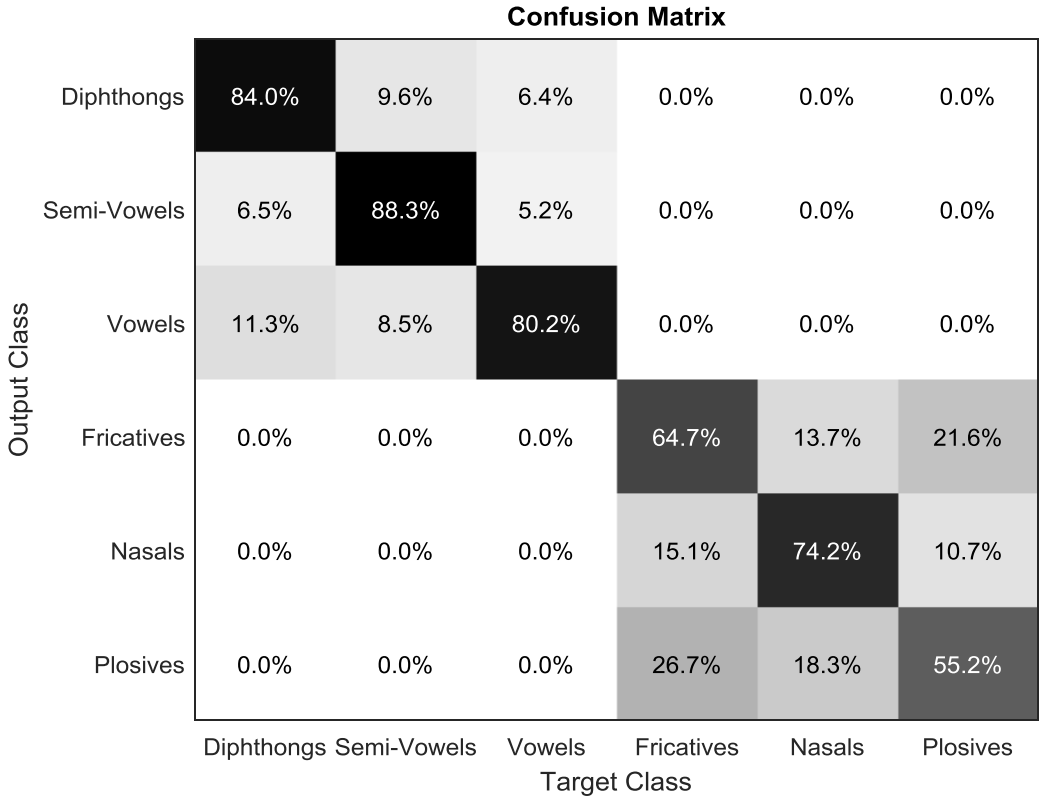
2.5.3- مرحلة التصنيف الثانية

بعد الانتهاء من مرحلة التصنيف الأولى والحصول على نوع الصوتيم (صامت أو صائت)، نقوم في هذه المرحلة بتصنيف كل نوع من أنواع الصوتيمات، وذلك إلى الصف الجزئي الموافق. تم اقتراح بينة لتصنيف كل من الصوامت والصوائت. يوضّح الجدول (3-19) البنية المقترحة لتصنيف كل من الصوامت والصوائت:

الجدول (3-19): بنية الشبكة العصبونيّة التلفيفية المقترحة لتصنيف كل من الصوامت والصوائت إلى الصفوف الجزئيّة الموافقة لها

نوع الطبقة	تابع التفعيل	ملاحظات
التفافية Conv1	RELU	32 مرشّح
BatchNormalization1		
التجميع Pool1		نواة بأبعاد 2×3 Mean Pooling
Dropout1		نسبة 15%
التفافية Conv2	RELU	64 مرشّح
BatchNormalization2		
التفافية Conv3	RELU	64 مرشّح
BatchNormalization3		
التجميع Pool2		نواة بأبعاد 2×3 Mean Pooling
Dropout2		نسبة 15%
التفافية Conv4	RELU	128 مرشّح
BatchNormalization4		
التجميع Pool3		نواة بأبعاد 2×2 Max Pooling
Dropout		نسبة 15%
التسوية Flatten		
طبقة الدخل للشبكة العصبونية كاملة الاتصال FC Input	RELU	عصبونات بطول طبقة التسوية
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden1	RELU	128 عصبون
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden2	RELU	32 عصبون
Dropout		نسبة 25%
طبقة الخرج للشبكة العصبونية كاملة الاتصال FC Out	Softmax	3 عصبون

بعد الانتهاء من تدريب الشبكة العصبونيّة لضبط المتحوّلات والأوزان الخاصّة بالشبكة، نقوم باختبار المصنّفين المقترحين على معطيات الاختبار. يوضح الشكل (3-13) رسم توضيحي لمصفوفة الالتباس الخاصّة بالمرحلة الثانية من التصنيف، والموافق لتصنيف كل نوع من الصوتيمات إلى الصف الجزئي الموافق:



الشكل (3-13): رسم توضيحي لمصفوفة الالتباس الناتجة عن تصنيف كل من الصوامت والصوائت إلى الصفوف الجزئيّة الموافقة

توضّح قيم القطر الرئيسي في مصفوفة الالتباس قدرتنا على تجزيئ الصوتيمات إلى صفوف جزئيّة بدقّة جيّدة نسبيّاً. ونلاحظ أن تصنيف الصوائت إلى صفوف جزئية أفضل منه في حالة تصنيف الصوامت إلى صفوف جزئيّة.

3.5.3- مرحلة التصنيف الثالثة

بعد الوصول إلى الصفوف الجزئيّة، يبقى أمامنا أن نقوم بتصنيف كل صف جزئي إلى الصوتيمات التي تنتمي إليه. قمنا باقتراح بنية شبكيّة واحدة لكل الصفوف الجزئيّة. يوضّح الجدول (3-20) البنية الشبكيّة المقترحة:

الجدول (3-20): الشبكة العصبونيّة التلفيفية المقترحة لتصنيف كل من الصفوف الجزئيّة إلى الصوتيمات الموافقة

نوع الطبقة	تابع التفعيل	ملاحظات
التفافية Conv1	RELU	32 مرشّح
BatchNormalization1		
التجميع Pool1		نواة بأبعاد 2×3 Mean Pooling
Dropout1		نسبة 15%
التفافية Conv2	RELU	64 مرشّح
BatchNormalization2		
التفافية Conv3	RELU	64 مرشّح
BatchNormalization3		
التجميع Pool2		نواة بأبعاد 2×3 Mean Pooling
Dropout2		نسبة 15%
التفافية Conv4	RELU	128 مرشّح
BatchNormalization4		
التفافية Conv5	RELU	128 مرشّح
BatchNormalization5		
التجميع Pool3		نواة بأبعاد 2×2 Mean Pooling
Dropout		نسبة 15%
التسوية Flatten		
طبقة الدخل للشبكة العصبونية كاملة الاتصال FC Input	RELU	عصبونات بطول طبقة التسوية
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden1	RELU	256 عصبون
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden2	RELU	128 عصبون
Dropout		نسبة 25%
طبقة مخفية للشبكة العصبونية كاملة الاتصال FC Hidden3	RELU	64 عصبون
Dropout		نسبة 25%
طبقة الخرج للشبكة العصبونية كاملة الاتصال FC Out	Softmax	عصبونات حسب عدد الصوتيمات التي تنتمي إلى الصف الجزئي

يكمن الاختلاف بين البنى الشبكيّة الخاصّة بكل صف جزئي بالأوزان والمتحوّلات، وذلك لأنّ كل شبكة عصبونية التفافيّة تمّ تدريبها على المعطيات التي تنتمي إلى الصف الجزئي الموافق فقط. يوضّح الجدول (3-21) نتائج تصنيف كل صف جزئي:

الجدول (3-21): نتيجة تصنيف كل من الصفوف الجزئيّة إلى الصوتيمات الموافقة لها

Phonemes classes	Accuracy
Plosives	55%
Nasals	74%
Fricatives	65%
Vowels	76%
Semi-Vowels	86%
Diphthongs	90%

بعد الانتهاء من مراحل التصنيف الثلاثة، يتم الآن حساب الدقّة على كامل الخوارزميّةٍ المقترحة، وذلك بناءً على المعادلة التالية:

$$\text{الدقة} = \frac{\boldsymbol{VC} \times [\boldsymbol{V} \times (\boldsymbol{VV} + \boldsymbol{VS} + \boldsymbol{VD}) + \boldsymbol{C} \times (\boldsymbol{CP} + \boldsymbol{CN} + \boldsymbol{CF})]}{\boldsymbol{NC}} \qquad (\mathbf{1-3})$$

حيث:

VC دقة التصنيف بين الصوتيمات الصوتية وغير الصوتية

V دقة تصنيف الصفوف الصوتية الثانوية

VV دقة تصنيف الصوتيمات Real-Vowels

VS دقة تصنيف الصوتيمات Semi-Vowels

VD دقة تصنيف الصوتيمات Diphthongs

C دقة تصنيف الصوتيمات غير الصوتية الثانوية

CP دقة تصنيف الصوتيمات Plosives

CN دقة تصنيف الصوتيمات Nasals

CF دقة تصنيف الصوتيمات Fricatives

NC عدد صفوف الصوتيمات الكلي

بالتعويض بالمعادلة (3-1) نجد:

$$\text{الدقة} = \frac{\mathbf{96} \times [\mathbf{84} \times (\mathbf{76} + \mathbf{86} + \mathbf{90}) + \mathbf{88} \times (\mathbf{55} + \mathbf{74} + \mathbf{65})]}{\mathbf{39}} = \mathbf{61\%}$$

وهكذا نكون قد حصلنا على دقّة في هذه الخوارزميّة مساوية لـ 61%.

6.3- النتيجة

نلاحظ من النتائج السابقة أن الطريقة المقترحة تعطي نتائج أفضل من الطريقة الأولى، حيث كانت نتيجة الطريقة المباشرة 57% بينما الطريقة المقترحة وصلت إلى 61%، وذلك لأننا في الطريقة المباشرة نعتمد على مرحلة تصنيف وحيدة باستخدام CNN، بينما في طريقتنا المقترحة هناك عدّة مراحل للتصنيف، **وفي كل مرحلة تصنيف يتم تدريب شبكة مختلفة على تصنيف جزئي، بمعنى أنّ الأوزان والمتحولات الخاصة بالشبكة تتناسب مع كل تصنيف جزئي. فمثلاً هنالك شبكة خاصة ومدرّبة على التصنيف بين الصوامت والصوائت.** ولكن تتفوق الطريقة المباشرة من حيث السرعة فهي بحاجة لمصنّف وحيد للحصول على التصنيف.

7.3- معلومات تقنيّة

1.7.3- اللغة البرمجية المستخدمة

قمنا باستخدام لغة بايثون Python والتي تعتبر من أكثر اللغات انتشارا ودعما للشبكات العصبونية، حيث تتميز هذه اللغة بالدعم الدائم والمتواصل من قبل المطّورين (مكاتب وصفوف). وتم الاعتماد على مكتبة Keras لبناء الشبكة العصبونية التلفيفية، وهي مكتبة مجانية مفتوحة المصدر، تتميز بسهولة بناء الشبكة العصبونيّة –حيث صممت بشكل أساسي لتكون صديقة للمستخدم– وباحتوائها على العديد من توابع التفعيل والمحسنات المختلفة. من أهم التوابع التي تم استخدامها:

- Sequential: لتعريف وبناء نماذج الشبكات العصبونية، وهو مكدس خطي من الطبقات.
- Compile: لتحديد وضبط إجرائية التعلم من تابع الخسارة والمحسن ومبدأ القياس.
- Dense: لتوصيف طبقة من طبقات الشبكة، حيث يتم تحديد عدد العصبونات وتابع التفعيل وغيرها.
- Convolutional: لتوصيف طبقة جداء التلاف.
- Pooling: لتوصيف طبقة التجميع.
- Flatten: للقيام بطبقة التسوية.
- Dropout: لتوصيف طبقة إطفاء بعض العصبونات عشوائيّاً.

وللتعامل مع هذه اللغة اخترنا برنامج يسمّى PyCharm وهو عبارة عن برنامج تطويري داعم للغة بايثون.

2.7.3- العتاد الصلب Hardware

لقياس أهميّة استخدام وحدة معالجة الرسوم في التدريب، قمنا بقياس زمن التدريب في حالة استخدام وحدة المعالجة العادية وحالة استخدام وحدة معالجة الرسوم. يوضّح الجدول (3-22) الأزمنة التي قسناها أثناء التدريب وذلك باستخدام عدّة أنواع مختلفة من العتاد:

الجدول (3-22): مقارنة العتاديات المستخدمة بالنسبة للزمن اللازم للتدريب

وحدة التدريب المستخدمة	الزمن اللازم للتدريب (الواحدة ساعة)
Core-I5 @2.20GHz	112
GeForce920M	28
Core-I7 @3.6GHz	50
GeForce GT 1030/4GB	12

ونلاحظ تفوّق واضح بالأداء عند استخدام وحدة معالجة الرسوم، وذلك لأنها مكوّنة من مجموعة كبيرة من المعالجات الصغيرة، والتي تعطي أداء عالي عند التعامل مع مصفوفات كبيرة من جهة، ومن جهة أخرى تتيح سرعة في التعامل مع العمليات التكرارية، حيث تعتبر هذه العمليات التكرارية الأساس في تدريب الشبكة العصبونيّة. تم الاعتماد في هذا العمل على حاسب يحوي وحدة معالجة الرسوم من نوع GeForce GT 1030/4GB وبمعالج Core-I7 وذاكرة RAM 16GB.

8.3- الخاتمة

قمنا في هذه الأطروحة بدراسة طرق التعرّف على الصوتيمات، واعتمدنا –نتيجة الدراسة– على استخدام التعلّم العميق، فانتقلنا إلى دراسة التعلّم العميق، ووقع اختيارنا –بناءً على النتائج والدراسات العلميّة– على الشبكة العصبونيّة التلفيفية. تمتاز هذه الشبكة بأعلى نتائج تصنيف عند التعامل مع دخل مصفوفاتي. تقدّم هذه الأطروحة دراسة كافية نظريّاً وعمليّاً عن الشبكة العصبونيّة التلفيفية، والتي تم استخدامها في هذا العمل لتصنيف الصوتيمات. بعد الاعتماد على مصنّف CNN قدّمنا دراسة عن تحويل MelSpectrogram الذي ينقل الإشارة الكلامية إلى المجال الترددي. يمتاز هذا التحويل بمحاكاة استجابة الأذن البشريّة للإشارة الكلاميّة. قمنا بعد ذلك بتنفيذ التعرّف على الصوتيمات باستخدام الطريقة التقليديّة (المباشرة) والتي تعتمد على مرحلة تصنيف وحيدة، وذلك لتكون لنا مرجع للمقارنة معه في طريقتنا المقترحة.

بعد الانتهاء من الطريقة التقليديّة، قدّمنا طريقتنا المقترحة، والتي تعتمد على تصنيف الصوتيمات عبر عدّة مراحل، وذلك بناءً على طريقة إصدار الصوتيم وانتماءه إلى الصفوف الجزئيّة. فيبدأ العمل بتصنيف الصوتيمات إلى صوامت وصوائت، ومن ثمّ تصنيف الصوامت والصوائت إلى الصفوف الجزئيّة، وهكذا حتى الوصول إلى الصوتيم المنطوق.

9.3- الآفاق المستقبلية

- ❖ من الممكن دمج أكثر من نوع من الشبكات العصبونيّة العميقة للحصول على أداء أفضل.
- ❖ العمل على تصنيف الصوتيمات في البداية إلى صوتيمات مجهورة وصوتيمات مهموسة باستخدام طرق معالجة الإشارة مثل كشف المرور بالصفر Zero-Crossing ومن ثمّ المتابعة بنفس طريق خوارزميتنا المقترحة.
- ❖ العمل على كشف تتالي من الصوتيمات لتكوين الكلمات ومحاولة الوصول إلى نظام ASR متكامل.
- ❖ الانتقال للعمل مع اللغة العربيّة وخاصّة أن التعامل مع الصوتيمات يعتبر مستقل عن اللغة.

الملاحق

الملحق آ

ضبط ثوابت الشبكة العصبونية في المرحلة الأولى

- **عدد طبقات التلاف والتجميع**

نتيجة تغير عدد الطبقات الأساسية على الدقّة

عدد الطبقات (طبقات تلاف وتجميع)	الدقّة%
2	88.2
3	93.1
4	92.2
5	91.2
6	91.5

نتيجة تغير عدد الطبقات الأساسية على الدقّة

عدد الطبقات التلاف	عدد الطبقات التجميع	الدقّة%
2	1	88.2
3	1	89.6
3	2	90.1
3	3	92.4
4	3	92.1
5	3	90.9
5	4	88.1
6	4	89.1
6	5	87.2

- **عدد المرشّحات في كل طبقة تلاف**

نتيجة تغير عدد المرشّحات في كل طبقة تلافيّة على الدقّة

عدد المرشّحات في كل طبقة	الدقّة%
[32,32,32]	89.4
[32,32,16]	90.3
[64,64,64]	89.8
[64,64,32]	91.1
[128 ,64,32]	91.5
[128,128, 64]	90.1

- **شكل طبقة التجميع**

الجدول نتيجة تغير متحوّلات طبقة التجميع على الدقّة

نواة التجميع	الخطوة	تابع التجميع	الدقّة%
2×3	2	MaxPooling	89.5
2×3	2	MeanPooling	91.1
2×2	2	MaxPooling	88.2
2×2	2	MeanPooling	89.2
2×3	3	MaxPooling	89.7
2×3	3	MeanPooling	90.2
2×2	3	MaxPooling	90.1
2×2	3	MeanPooling	90.8

- **إضافة طبقات Dropout**

الجدول نتيجة إضافة طبقة Dropout على الدقّة

طبقة Dropout	الدقّة%
غير موجودة	88
عدّة طبقات (بعد طبقات التلاف والتجميع وبعد كل طبقة من طبقات FC)	90.2

- **طبقة FC**

نتيجة تغير عدد الطبقات المخفيّة وعدد العصبونات في كل منها على الدقّة

عدد الطبقات المخفية	عدد العصبونات في كل طبقة	الدقّة
1	[128]	93.2
1	[256]	93.4
1	[512]	93.1
2	[32 32]	94
2	[64 64]	94.1
2	[64 32]	94.5
2	[128 32]	95.1
2	[128 64]	94.8
3	[64 64 64]	92.2
3	[128 64 32]	92.1
3	[128 128 128]	92.4

الملحق ب

ضبط ثوابت الشبكة العصبونية في المرحلة الثانية

- **عدد طبقات التلاف والتجميع**

نتيجة تغير عدد الطبقات الأساسية على الدقّة

عدد الطبقات (طبقات تلاف وتجميع)	الدقّة%
2	70
3	71.6
4	72.2
5	72
6	71.3

نتيجة تغير عدد الطبقات التلفيفية وطبقات التجميع على الدقّة

عدد الطبقات التلاف	عدد الطبقات التجميع	الدقّة%
2	1	69.2
3	1	69.8
3	2	70.2
3	3	69.8
4	3	71.5
5	3	71.1
5	4	70.5
6	4	70.1
6	5	69.9

- **عدد المرشّحات في كل طبقة تلاف**

نتيجة تغير عدد المرشّحات في كل طبقة تلافيّة على الدقّة

عدد المرشّحات في كل طبقة	الدقّة%
[32,32,32,32]	68.2
[32,32,32,16]	68.4
[64,64,64,64]	69.2
[128,64,32,32]	69.1
[128 ,64,64,32]	70.2
[128,128, 64,64]	69.7

- **شكل طبقة التجميع**

نتيجة تغير متحوّلات طبقة التجميع على الدقّة

نواة التجميع	الخطوة	تابع التجميع	الدقّة%
2×3	2	MaxPooling	69.5
2×3	2	MeanPooling	71.5
2×2	2	MaxPooling	66.5
2×2	2	MeanPooling	68.3
2×3	3	MaxPooling	68.4
2×3	3	MeanPooling	69.4
2×2	3	MaxPooling	65.3
2×2	3	MeanPooling	66.2

- **إضافة طبقات Dropout**

نتيجة إضافة طبقة Dropout على الدقّة

طبقة Dropout	الدقّة%
غير موجودة	64.2
عدّة طبقات (بعد طبقات التلاف والتجميع وبعد كل طبقة من طبقات FC)	69.7

- **طبقة FC**

نتيجة تغير عدد الطبقات المخفيّة وعدد العصبونات في كل منها على الدقّة

عدد الطبقات المخفية	عدد العصبونات في كل طبقة	الدقّة
1	[128]	60.2
1	[256]	60.1
1	[512]	61.5
2	[32 32]	65.3
2	[64 64]	66.5
2	[64 32]	64.3
2	[128 32]	69.5
2	[128 64]	68.4
3	[64 64 64]	67.2
3	[128 64 32]	66.5
3	[128 128 128]	65.3

المراجع

[1] Pradeep R, K. Sreenivasa Rao," Deep Neural Networks for Kannada Phoneme Recognition", 2016 IEEE.

[2] Elyes Zarrouk, Yassine Benayed, "Hybrid SVM/HMM Model for the Arab Phonemes Recognition", The International Arab Journal of Information Technology, Vol. 13, No. 5, September 2016.

[3] Ying Zhang, Mohammad Pezeshki, Phil´emon Brakel, Saizheng Zhang, C´esar Laurent Yoshua Bengio, Aaron Courville," Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks" arXiv:1701.02720v1 [cs.CL] 10 Jan 2017.

[4] Asadolahzade Kermanshahi, M., and M. M. Homayounpour. "Improving Phoneme Sequence Recognition using Phoneme Duration Information in DNN-HSMM" Journal of AI and Data Mining 7.1 (2019): 137-147.

[5] Chiu, Chung-Cheng, et al. "State-of-the-art speech recognition with sequence-to-sequence models." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[6] Zheng, Xin, et al. "Learning dynamic features with neural networks for phoneme recognition." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014

[7] Nagamine, Tasha, Michael L. Seltzer, and Nima Mesgarani. "Exploring how deep neural networks form phonemic categories."Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[8] Fauziya, Farheen, and Geeta Nijhawan. "A Comparative study of phoneme recognition using GMM-HMM and ANN based acoustic modeling." International Journal of Computer Applications 98.6 (2014).

[9] Palaz, Dimitri, Ronan Collobert, and Mathew Magimai Doss. "End-to-end phoneme sequence recognition using convolutional neural networks." arXiv preprint arXiv:1312.2137 (2013).

[10] Ahmad, Khan Suhail, et al. "A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network." 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR). IEEE, 2015.

[11] Yousafzai, Jibran, et al. "Combined features and kernel design for noise robust phoneme classification using support vector machines." IEEE Transactions on Audio, Speech, and Language Processing 19.5 (2010): 1396-1407.

[12] Hamooni, Hossein, Abdullah Mueen, and Amy Neel. "Phoneme sequence recognition via DTW-based classification." Knowledge and Information Systems 48.2 (2016): 253-275.

[13] Glackin, Cornelius, et al. "Convolutional Neural Networks for Phoneme Recognition." ICPRAM. 2018.

[14] Mohamed, Abdel-rahman, George E. Dahl, and Geoffrey Hinton. "Acoustic modeling using deep belief networks." IEEE transactions on audio, speech, and language processing 20.1 (2011): 14-22.

[15] عفاف الشلبي، د.أميمة الدكاك و د.ندى غنيم. أطروحة دكتوراه بعنوان "نحو نظام لتركيب الكلام باللغة العربية من نصوص في المعهد العالي للعلوم التطبيقة والتكنولوجيا باستعمال الضم لأنصاف مقاطع صوتيّة وتنغيم طبيعي". 2018

[16] Reynolds, T. Jeff, and Christos A. Antoniou. "Experiments in speech recognition using a modular MLP architecture for acoustic modelling." Information Sciences 156.1-2 (2003): 39-54.

[17] Feng, Ling. Speaker recognition. MS thesis. Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2004.

[18] Jagadeeshchandra, Anupama Sira. Speaker Identification and Voice Impairments detection. Diss. Siauliai University.

[19] G.Suvarna Kumar, K.A.Prasad Raju, Dr.Mohan Rao Cpvnj, P.Satheesh, "Speaker Recognition Using GMM," International Journal of Engineering Science and Technology, vol. 2(6), pp. 2428-2436, 2010.

[20] Do, Ngoc. "Neural networks for automatic speaker, language, and sex identification." (2016).

[21] Su, Yu, et al. "Environment sound classification using a two-stream cnn based on decision-level fusion Sensors 19.7 (2019): 1733.

[22] Choi, Keunwoo, George Fazekas, and Mark Sandler. "Automatic tagging using deep convolutional neural networks." arXiv preprint arXiv:1606.00298 (2016).

[23] H. of Lords Select Committee on Artificial Intelligence, AI in the UK: ready, willing and able?, House of Lords. 36 (2018).

[24] Ian Goodfellow, Y. Bengio, A. Courville, Deep learning, Nat. Methods. 13 (2017)
35. doi:10.1038/nmeth.3707.

[25] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. Neural Computation 1, 541-551,1989.

[26] Hubel, David H., and Torsten N.Wiesel. "Receptive fields of single neurones in the cat's striate cortex." The Journal of physiology 148.3: 574-591. 1959.

[27] Krizhevsky, Sutskever, & Hinton, ImageNet Classification with Deep Convolutional Neural Networks.Advances In Neural Information Processing Systems, 1–9. 2012.

[28] M. HODOSH, P. YOUNG AND J. HOCKENMAIER, "FRAMING IMAGE DESCRIPTION AS A RANKING TASK:DATA, MODELS AND EVALUATION METRICS," JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH , NO. 47, PP. 853-899, 2013.

[29] Talathi, Sachin S., and Aniket Vartak. "Improving performance of recurrent neural network with ReLU nonlinearity." arXiv preprint arXiv:1511.03771 (2015).

[30] Le, Hai-Son, et al. "Structured output layer neural network language model." 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011.

[31] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015).

الملخص

أخذت الشبكات العصبونية العميقة في السنوات القليلة الماضية مسألة التعرّف الآلي على الصوت إلى مستويات جديدة من الدقة[1] . حيث حازت على أعلى نسب للتعرّف، سواء على الكلمات بشكل مفرد أو على الصوتيمات. تمثل مسألة التعرّف على الصوتيمات المرحلة الأولى من مراحل التعرّف في أنظمة التعرّف الآلي على الكلام. نقدم في هذا البحث التعرّف على الصوتيمات اعتماداً على الشبكات العصبونية العميقة باستخدام الشبكات العصبونية التلفيفية Convolutional neural network 'CNN'. حيث نقدم طريقتين للتعرّف الطريقة الأولى المباشرة عن طريق التعرّف على الصوتيمات بمرحلة تصنيف وحيدة، وذلك بالحصول على نوع الصوتيم مباشرة عن طريق الدخل. أمّا الطريقة الثانية المقترحة تتم عن طريق عدة مراحل للتصنيف وذلك بأخذ طريقة إصدار الصوتيمات وصفوفها بعين الاعتبار (صائت vowels ونصف صائت semi-vowels وصوامت consonants و...).

واعتمدنا في الطريقتين على تحويل MelSpectrogram، حيث يجري تحويل الإشارة الصوتية إلى مصفوفة ثنائية البعد ضمن الفضاء الترددي، ومن ثم يتم إدخال هذه المصفوفة كدخل للشبكة العصبونية العميقة.

قمنا باختبار المصنف المقترح على قاعدة المعطيات TIMIT، وحصلنا على دقة 57% في الطريقة المباشرة، وحصلنا على دقة أعلى باستخدام طريقتنا المقترحة 61%.

Abstract

In the last few years, deep neural networks have taken the problem of automated voice recognition to a completely new level of accuracy. Where it provided the highest recognition rates, whether on words or on phonemes.

Voice recognition problem represents the first phase of automated speech recognition systems. In this research, we introduce the recognition of phonemes based on deep neural networks using the Convolutional neural network 'CNN'.

We will discuss two methods of recognition, the direct method by recognizing the phonemes using a single classification phase by obtaining the correct phonemes directly through the input. The second proposed method uses several phases of classification by taking into account the types of phonemes and their classes (vowels, semi-vowels, explosive, etc.). In both methods, we rely on the Mel Spectrogram transform, where the acoustic signal converted into a two-dimensional matrix within the frequency domain, this matrix then inserted as the input of the deep neural network.

We tested the proposed classifier on TIMIT database, obtained 57% accuracy in the direct method, and a higher accuracy of 61% using our proposed method.