Syrian Arab Republic
Higher Institute for Applied Sciences and Technology
Informatics Department

A thesis submitted for
Master degree in Big Data systems

Mispronunciation detection in Arabic language using Deep Neural Networks

Submitted by

Eng. Elham Almfashi

Supervised by

Dr. Oumayma Dakkak **Dr. Nada Ghneim**

2021

الجمهورية العربية السورية

المعهد العالي للعلوم التطبيقية والتكنولوجيا

قسم المعلوميات

أعدت هذه الأطروحة لنيل

درجة الماجستير في نظم المعطيات الكبيرة

كشف أخطاء النطق في النظم المساعدة على تعليم اللغة العربية باستعمال شبكات التعلم العميق DNN

إعداد

م. إلهام المفشّي

إشراف

د. أميمة الدكاك

د. ندى غنيم

2021

المعهد العالي للعلوم التطبيقية والتكنولوجيا

Higher Institute for Applied Sciences & Technology

المعهد العالي للعلوم التطبيقية والتكنولوجيا مؤسسة حكومية للتعليم العالي أحدث بموجب المرسوم التشريعي رقم /٢٤/ لعام ١٩٨٣، وذلك بهدف إعداد كوادر علمية متميزة من مهندسين وباحثين للإسهام الفاعل في عملية التطوير العلمي والتنمية في الجمهورية العربية السورية.

يمنح المعهد العالي درجة الإجازة في الهندسة في الاتصالات والمعلوماتية والنظم الإلكترونية والميكاترونيكس وعلوم وهندسة المواد وهندسة الطيران. يقبل المعهد العالي دراسة هذه الاختصاصات شريحة منتقاة من المتفوقين في الشهادة الثانوية من الفرع العلمي. يتيح المعهد العالي أيضاً برامج ماجستير أكاديمي في نظم الاتصالات وفي التحكم والروبوتيك وفي نظم المعطيات الكبيرة ونظم المعلومات ودعم القرار وفي علوم وهندسة المواد وعلوم وهندسة البصريات. وأخيراً، يمنح المعهد العالي درجة الدكتوراه في الاتصالات والمعلوماتية ونظم التحكم والفيزياء التطبيقية. تُحدث في المعهد العالي اختصاصات جديدة بحسب متطلبات سوق العمل وتوجهات البحث والتطوير المحلية والعالمية.

يمارس المعهد العالي عبر جهود أطره وفعالياته العلمية المختلفة نشاطاً حثيثاً في البحث والتطوير، إذ ينفذ مشاريع ودراسات واستشارات متنوعة لصالح الجهات العامة والخاصة في القطر، كما يتعاون مع جهات خارج القطر في بعض المشاريع البحثية والتطويرية. يسعى المعهد أيضاً، عبر دورات تدريبية نظرية وعملية متاحة للقطاعين العام والخاص وللأفراد، إلى إفادة أوسع فئة من المهتمين من إمكانيات العلمية ومختبراته. واستكمالاً لدوره الرائد في مجال التعليم ونشر العلم، ينشر المعهد العالي كتباً علمية عالية المستوى من نتاج أطره، منها ما هو تدريسي يوافق المناهج في المعهد العالي ويفيد شريحة واسعة من الطلاب الجامعيين عموماً، ومنها ما هو علمي ثقافي. يتيح المعهد العالي بعضاً من منشوراته على موقعه على الشابكة، كما يتيح إمكانية الاطلاع على رسائل الماجستير والدكتوراه المنفذة في المعهد العالي وعلى بعض منشورات طلابه وأطره من المقالات العلمية.

المعهد العالي للعلوم التطبيقية والتكنولوجيا، الجمهورية العربية السورية، دمشق، ص.ب ٣١٩٨٣

Higher Institute for Applied Sciences & Technology – HIAST

P. O. Box 31983, Damascus, Syrian Arab Republic

هاتف ٠٠٩٦٣١١٥١٢٣٨١٩ – فاكس ٠٠٩٦٣١١٥١٤٠٧٦١

البريد الإلكتروني contact@hiast.edu.sy

الموقع الإلكتروني www.hiast.edu.sy

شكر

﴿ رَبِّ أَوْزِعْنِي أَنْ أَشْكُرَ نِعْمَتَكَ الَّتِي أَنْعَمْتَ عَلَيَّ وَعَلَىٰ وَالِدَيَّ وَأَنْ أَعْمَلَ صَالِحًا تَرْضَاهُ ﴾

أحمد الله تعالى أن وفقني لهذا العمل ويسّر أمري فيه ومنحني العزم والصبر على إنجازه، وما توفيقي إلا به سبحانه وتعالى.

وكما علمنا النبي صلى الله عليه وسلم في هديه (من لا يشكر الناس لا يشكر الله)، فإني أتوجه بجزيل الشكر والامتنان والتقدير لكل من أفاض عليّ جهده ووقته ومنحني معونته من لحظة بذر فكرة العمل وحتى حصاد نتائجها. وأخص بالشكر مشرفتيَّ الدكتورة أميمة الدكاك والدكتورة ندى غنيم لإرشادهما لي في جميع خطوات البحث وحثي الدائم على العمل والمتابعة، وتقديم الملاحظات القّيمة طيلة فترة العمل. كما أتوجه لأعضاء الكادر التدريسي في المعهد العالي بجزيل الشكر لما قدموه لنا خلال فترة الدراسة من علوم ومعارف. وأقدم تحية خاصة للدكتور ياسر رحال والدكتورة غيداء ربداوي للتشجيع والدعم النفسي خلال فترة العمل على البحث.

وأتوجه بالشكر الجزيل لسكرتيرة قسم المعلوميات (آ. ماري) لما قدمته لنا من مساعدة وتسهيلات خلال فترة الماجستير.

إهداء

أتقدم بخالص امتناني ومحبتي وجميل دعائي وشكري لمنبعَي الحب والعطاء (أبي وأمي) وأهديهما هذا النجاح مكلّلاً بحبهما ورضاهما.

وفي سطوري أخطّ تحية معطرة مغلّفة بأسمى مشاعر الحب لقبس النور والعلم في حياتي، نبض الحياة لروحي (معلّمتي آ. هدى).

وأهدي فرحي وحبي وشكري لأختي التي لم تلدها أمي، الساكنة في سويداء قلبي.

كما أبثُّ لصديقتَي الدرب، السند والكتف في رحلة البحث (سنا ومروة) شكراً لا تسَعُه السطور.

وختاماً أشمل بالشكر والعرفان إخوتي وصديقاتي وكل من دعا لي بالتوفيق وكان سنداً لي في طريقي ومسيرة بحثي.

الملخص

يعد كشف النطق الخاطئ من الأمور الهامة في نظم تعلّم اللغات بمساعدة الحاسوب Computer-Aided Language Learning (CALL)، حيث يساعد تحديدُ أماكن الخطأ في النطق متعلّم اللغة في الحصول على تقييم دقيق لصحة اللفظ. وقد لاقت هذه النظم اهتماماً كبيراً لأنها تمكن متعلّمي اللغة من تحسين قدراتهم اللغوية دون الحاجة للتواصل مع المختصين اللغويين بشكل مباشر، وذلك بالاستفادة من وسائل التعلّم الحديثة والتقنيات المتطورة. يهدف هذا البحث لإيجاد المنهجيات المناسبة في بناء نظام يساعد متعلّم اللغة على معرفة موضع الخطأ في نطقه وكيفية تصحيحه، وذلك بدراسة إمكانية تصنيف الصوتيمات phonemes وتمييزها آلياً وخاصة المتشابه والمشترك منها في مخرج النطق وبعض صفات الحروف بالاعتماد على شبكات التعلم العميق Deep neural networks، إضافة لدراسة إمكانية تمييز واصفات النطق الكلامية speech attributes بهدف استخدامها في كشف خطأ النطق وتحديد نوعه كونها تعطي صورة عن صوت الحرف حسب مكان خروجه وصفاته. اقترحنا في البحث استخدام التعلم المتعدد المهام (multitask learning) للقيام بمهمة تمييز الواصفات. قمنا باختبار أداء المنهجية المقترحة والمقارنة مع منهجيات أخرى، وحصلنا على نتائج أفضل بالنسبة لتعرف واصفات النطق. إن نسبة التعرف الجيدة للواصفات تمكن من توظيف هذا النموذج في أنظمة تعليم اللغات وتفتح المجال لاستخدام هذه الواصفات في تطبيقات أخرى كتركيب الكلام وتحويل الصوت لنص مكتوب. ركزنا في البحث على اللغة العربية بهدف تقليص الهوة البحثية الموجودة بين التقانات اللغوية الداعمة للّغة العربية ومثيلاتها في اللغات العالمية والتي حققت تقدماً كبيراً في العديد من المجالات.

Abstract

Mispronunciation Detecting is an important issue in computer-aided language learning (CALL) systems, where detecting errors in pronunciation helps the language learner to obtain an accurate pronunciation correctness assessment. these systems have received great attention because they give language learners the possibility to improve their language proficiency without the need for direct communication with language specialists, by making use of modern learning methods and advanced technologies. This research aims to find appropriate methodologies to build a system that helps language learners to detect mispronunciations and provide a corrective feedback, by studying the possibility of classifying phonemes and distinguishing them automatically, especially the similar and common ones in place and manner of articulation based on deep learning networks. In addition, this research studies the possibility of distinguishing speech attributes in order to use them in detecting pronunciation errors' position and type, as these attributes describe the phoneme according to its place and manner of articulation. we proposed the use of multitask learning to perform the attribute detection task. The performance of our proposed methodology was tested and compared with other methodologies, and better results were obtained for detecting the speech attributes. These good results enable the employment of this model in language learning systems and open the way to use these attributes in other applications such as speech synthesis and speech-to-text systems. The focus of the research was on Arabic language, aiming to reduce the exist research gap between the linguistic technologies that support Arabic language and its equivalent in international languages, which have achieved great progress in many fields.

الفهرس

قائمة الأشكال viii

قائمة الجداول ix

جدول اختصارات ومصطلحات..... x

الفصل الأول: مقدمة 1

1.1. دوافع البحث 1

2.1. إشكالية البحث 2

3.1. فكرة الحل المقترح..... 2

4.1. مساهمات البحث 2

5.1. مخطط البحث..... 3

الفصل الثاني: الدراسة النظرية 4

1.2. علم الأصوات 4

1.1.2. مقدمة..... 4

2.1.2. الأصوات في اللغة العربية..... 5

3.1.2. معالجة الإشارة الصوتية 11

2.2. شبكات التعلم العميق..... 14

1.2.2. تعريف شبكة التعلم العميق..... 15

2.2.2. أساسيات في التعلم العميق 16

3.2.2. اعتبارات عملية 17

4.2.2. أصناف شبكات التعلم العميق..... 19

5.2.2. شبكات التعلم العميق Deep Neural Networks (DNNs)..... 20

6.2.2. توابع التفعيل Activation Functions..... 22

7.2.2. الذاكرة طويلة قصيرة الأمد Long Short Term Memory (LSTM)..... 24

3.2. خاتمة 26

الفصل الثالث: الدراسة المرجعية 27

1.3. مقدمة 27

2.3. المنهجيات العامة المتبعة..... 28

1.2.3. المنهجيات المعتمدة على درجة الثقة (Confidence score based)..... 28

2.2.3. المنهجيات المعتمدة على القواعد (Rule based)..... 29

3.2.3. المنهجيات المعتمدة على المصنفات (Classifier based)..... 29

4.2.3. المنهجيات المعتمدة على التعلم العميق (Deep Neural network based)..... 30

3.3. الأعمال السابقة..... 30

4.3. خاتمة 34

الفصل الرابع: المقاربة المقترحة ومنهجية العمل..... 36
1.4. مقدمة 36
2.4. مخطط عام..... 36
3.4. تحضير المعطيات ومعالجتها 37
4.4. مرحلة التعرف وتدريب شبكات التعلم..... 40
1.4.4. تعرّف الصوتيمات 41
2.4.4. تعرّف واصفات الكلام..... 43
5.4. خاتمة 45
الفصل الخامس: الاختبارات والنتائج..... 46
1.5. المعطيات المستخدمة..... 46
.1.1.5 مجموعة معطيات النطق بالعربية Arabic Speech Corpus..... 46
2.1.5. مجموعة المعطيات TIMIT (Texas Instruments Massachusetts Institute of Technology)..... 49
.3.1.5 مجموعة معطيات الصوتيات العربية KACST Arabic Phonetic Database (KAPD)..... 50
.2.5 مقاييس التقييم..... 51
3.5. تعرّف الصوتيمات وواصفات الكلام في مجموعة معطيات KAPD 52
.1.3.5 تعرف الصوتيمات في KAPD 52
.2.3.5 تعرف واصفات الكلام في KAPD..... 57
.4.5 تعرّف الصوتيمات وواصفات الكلام في مجموعة معطيات النطق بالعربية MSA 60
1.4.5. تعرف الصوتيمات في مجموعة معطيات النطق بالعربية..... 60
.2.4.5 تعرف واصفات الكلام في مجموعة معطيات النطق بالعربية 62
.5.5 تعرّف الصوتيمات وواصفات الكلام في مجموعة معطيات TIMIT 65
1.5.5. تعرف الصوتيمات في مجموعة معطيات TIMIT 65
2.5.5. تعرف واصفات الكلام في مجموعة معطيات TIMIT..... 68
الفصل السادس: الخاتمة والآفاق المستقبلية..... 69
1.6. خاتمة 69
2.6. الآفاق المستقبلية..... 70
المراجع..... 71

قائمة الأشكال

الشكل 1- مخارج الحروف الرئيسية.....6
الشكل 2- مجموعة مرشحات مثلثية Triangular filterbank.....13
الشكل 3- استخلاص شعاع سمات MFCC من الإشارة الصوتية 13
الشكل 4- مخطط فين للعلاقة بين تخصصات الذكاء الصنعي المختلفة.....15
الشكل 5- شبكة تعلم عميق بطبقة دخل وطبقة خرج وطبقتين خفيتين 21
الشكل 6- تابع التفعيل ReLU.....22
الشكل 7- تابع التفعيل Sigmoid 23
الشكل 8- تابع التفعيل Tanh 23
الشكل 9- بنية خلية ذاكرة LSTM.....25
الشكل 10- أنواع أخطاء النّطق 27
الشكل 11- نموذج العمل المقترح.....36
الشكل 12- التشارك القاسي للموسطات في التعلم المتعدد المهام.....44
الشكل 13- التشارك اللّين للموسطات في التعلم المتعدد المهام 44
الشكل 14- البنية العامة لشبكة تعرف واصفات الكلام 45
الشكل 15- قياس ضبط الإطار لشبكة LSTM لبيانات التدريب والاختبار والتحقق بدلالة عدد دورات التدريب 53
الشكل 16- ضبط التصنيف للصوتيم لشبكة LSTM مع تغيير عدد معاملات MFCC بدلالة عدد دورات التدريب 54
الشكل 17- مقارنة مقاييس التقييم مع تغيير بنية شبكة التعلم (عدد الطبقات وحجم كل منها) 54
الشكل 18- ضبط التصنيف بدلالة عدد دورات التدريب من أجل قيم مختلفة لمعامل التعلم.....55
الشكل 19- مصفوفة الالتباس في مجموعة معطيات KAPD 56
الشكل 20 - نتائج تصنيف واصفات الكلام على مستوى الصوتيم باستخدام مقياس F1 59
الشكل 21- ضبط التصنيف للصوتيم في مجموعة معطيات النطق بالعربية باستخدام عدد طبقات مختلفة لشبكةLSTM.....60
الشكل 22- مصفوفة الالتباس لصوتيمات مجموعة معطيات اللغة العربية.....62
الشكل 23- ضبط التصنيف في مجموعة معطيات TIMIT من أجل بنى مختلفة لشبكة LSTM ...66

قائمة الجداول

الجدول 1 أصوات اللغة العربية الفصحى 10
الجدول 2 مخارج الحروف لأصوات اللغة العربية 10
الجدول 3 صفات الحروف لأصوات اللغة العربية 11
الجدول4 مقارنة بين الأعمال السابقة 35
الجدول 5 واصفات الكلام في اللغة العربية 38
الجدول 6 واصفات الكلام في اللغة الإنجليزية 39
الجدول 7 مجموعة صوتيمات مجموعة معطيات النطق بالعربية 47
الجدول 8 التقابل بين 82 إلى 38 صوتيم في مجموعة صوتيمات مجموعة معطيات النطق بالعربية.48
الجدول 9 صوتيمات مجموعة معطيات النطق بالعربية بعد التقليص إلى 38 صوتيم 48
الجدول 10 صوتيمات مجموعة معطيات TIMIT بعد التقليص إلى 39 صوتيم 49
الجدول 11 صوتيمات مجموعة المعطيات KAPD 50
الجدول 12 نتائج تعرف الصوتيمات في KAPD باستخدام شبكة LSTM مع عدد إطارات مجاورة مختلفة 52
الجدول 13 توزع صوتيمات KAPD في بيانات التدريب والاختبار 53
الجدول 14 مقاييس التقييم لتعرّف الصوتيمات في KAPD باستخدام شبكة LSTM 55
الجدول 15 معدل خطأ الصوتيم في KAPD بالمقارنة مع منهجيات أخرى 55
الجدول 16 مجال قيم التجريب لموسطات شبكة DNN.....57
الجدول 17 نتائج مقياس F1 لشبكة DNN لتصنيف واصفات الكلام في KAPD 58
الجدول 18 نتائج تقييم بيانات اختبار مجموعة معطيات النطق بالعربية لشبكة LSTM بثلاث طبقات خفية.....60
الجدول 19 مثال على وسم كلمات تحوي ياء مشددة في مجموعة معطيات النطق بالعربية 61
الجدول 20 موسطات شبكةDNN لتعرّف واصفات الكلام في MSA.....63
الجدول 21 نتائج تصنيف شبكةDNN لواصفات الكلام في مجموعة معطيات اللغة العربية.....64
الجدول 22 مقاييس التقييم لتعرّف الصوتيمات في TIMIT باستخدام شبكة LSTM 66
الجدول 23 مقارنة خطأ تعرّف الصوتيم في TIMIT مع نماذج مختلفة 66
الجدول 24 نتائج ضبط تصنيف شبكة DNN لواصفات الكلام في TIMIT 68

جدول اختصارات ومصطلحات

الاختصار	اللغة الإنجليزية	اللغة العربية
IPA	International phonetic alphabet	الأبجدية الصوتية العالمية
	Feature extraction	استخلاص السمات
	Feature normalization	استنظام (تسوية) السمات
	Cross-entropy	الإنتروبية التقاطعية
	Data	بيانات / معطيات
MLPs	Multilayer perceptrons	البيرسيبترون متعدد الطبقات
	Activation function	تابع التفعيل
	Loss function	تابع الخسارة
	Objective function	تابع الهدف
FFT	Fast Fourier transform	تحويل فورييه السريع
DFT	Discrete Fourier transform	تحويل فورييه المتقطع
IDFT	Inverse Discrete Fourier Transform	تحويل فورييه المتقطع العكسي
CAPT	Computer aided pronunciation training	التدرب على النطق بمساعدة الحاسوب
	Gradient descent	التدرج المنحدر
SGD	Stochastic gradient descent	التدرج المنحدر العشوائي
ASR	Automatic speech recognition	تعرف الكلام الآلي
	Machine learning	تعلّم الآلة
	Deep learning	التعلم العميق
CALL	Computer-aided language learning	تعلم اللغات بمساعدة الحاسوب
	Multi-task learning	التعلّم المتعدد المهام
	Transfer learning	التعلّم المنقول
	Optimization algorithm	خوارزمية التحسين
	Precision	الدقة
LSTM	Long short term memory	الذاكرة طويلة قصيرة الأمد
	Artificial intellegence	الذكاء الصنعي
	Features	السمات
DNNs	Deep neural networks	شبكات التعلم العميق
	Neural netowrks	الشبكات العصبونية

RNNs	Recurrent neural networks	الشبكات العصبونية التكرارية
CNNs	Convolutional neural networks	الشبكات العصبونية التلافيفية
DBNs	Deep belief networks	شبكات المعتقدات العميقة
GOP	Goodness of Pronunciation	صحة النّطق
	Consonants	الصوامت
	Vowels	الصوائت
	Phoneme	الصوتيم
	Accuracy	الضبط
	Neuron / Node	عصبون / عقدة
	Phonetics	علم الأصوات اللغوية
	Phonology	علم وظائف الأصوات اللغوية
	Manner of Articulation	كيفية النطق
KAPD	Kacst arabic phonetic database	مجموعة المعطيات الصوتية KAPD
TIMIT	Texas Instruments Massachusetts Institute of Technology	مجموعة المعطيات الصوتية TIMIT
	Dataset	مجموعة معطيات
	Place of Articulation	مخرج النطق (مخرج الحرف)
MFCCs	Mel frequency cepstral coefficients	معاملات ميل التردّدية
LR	Learning rate	معدل التعلّم
FRR	False rejection rate	معدل الرفض الخاطئ
FAR	False acceptance rate	معدل القبول الخاطئ
PER	Phoneme error rate	معدل خطأ الصوتيم
	Syllables	المقاطع الصوتية
	International phonetic association	المنظمة العالمية للصوتيات
	Pronunciation	النّطق
GMM	Gaussian mixture model	نموذج المزج الغاوسي
HMM	Hidden markov model	نموذج ماركوف المخفي
	Attribute	واصفة
GPU	Graphics processing unit	وحدة المعالجة الرسومية
	Label	وسم

الفصل الأول: مقدمة

يعد كشف النطق الخاطئ من الأمور الهامة في نظم تعلم اللغات بمساعدة الحاسوب Computer-Aided Language Learning (CALL)، حيث يساعد تحديدُ أماكن الخطأ في النطق متعلمَ اللّغة في الحصول على تقييم دقيق لصحة اللفظ. وقد لاقت هذه النظم اهتماماً كبيراً لأنها توفر لمتعلمي اللغة القدرة على تحسين قدراتهم اللغوية دون الحاجة للتواصل مع المختصين اللغويين بشكل مباشر وذلك بالاستفادة من وسائل التعلم الحديثة والتقنيات المتطورة.

1.1. دوافع البحث

تركز البحث في أحد جوانب نظم تعلم اللغات بمساعدة الحاسوب على كشف النطق الخاطئ وتصحيحه Computer Aided Pronunciation Training (CAPT) بهدف تحسين النطق عند المتحدثين غير الأصليين للغة وذلك بتحديد مواضع الخطأ في الكلام والتدريب على النطق الصحيح بطرق وآليات مختلفة [1]، وقد لاقى هذا المجال اهتماماً كبيراً من قبل الباحثين نظراً لضرورته واستخدامه في تطبيقات عديدة، إضافة للتقدم الملحوظ في تقنيات الذكاء الصنعي وتعلّم الآلة. تهدف هذه الأنظمة إلى استخدام الموارد الحاسوبية والتقنيات الحديثة في تسهيل عملية تعلّم اللغة وحل مشكلة صعوبة التواصل بشكل دائم بين مدرس اللغة ومتعلّمها. أُجريت العديد من الدراسات في هذا المجال من أجل مختلف اللغات (الإنجليزية [2]، والصينية [3]، والألمانية، واليابانية، والعربية [4]...) لكشف النطق الخاطئ وتصحيحه، وتم العمل على التحقق من النطق وفق عدة مستويات بدءاً من مستوى المتحدث speaker، بهدف تقييم طلاقة الكلام عند النطق واستخدامه في اختبارات الكفاءة اللغوية المنطوقة، وانتهاءً بمستوى الصوتيم phoneme على مستوى كل وحدة صوتية في الكلام. يقدم مستوى الصوتيم معلومات أكثر دقة عن مكان ونوع الخطأ الذي ارتكبه المستخدم[5]، وهذا يجعل عملية التعلّم أكثر فاعلية وأعلى جودة وكفاءة. على الرغم من ذلك لا يزال العمل على هذا المستوى للحصول على نظام دقيق جداً لكشف الخطأ يشكل تحدياً بالنسبة للباحثين في المجال [6].

2.1. إشكالية البحث

يركز هذا البحث على دراسة إمكانية تصنيف الصوتيمات وتمييزها آلياً وخاصة المتشابه والمشترك منها في مخرج الصوت وبعض صفات الحروف، إضافة لدراسة أثر استخدام واصفات النطق الكلامية للمساعدة في بناء نظام يساعد متعلّم اللغة على كشف أخطاء النطق وتصحيحها. فعلى الرغم من وجود دراسات سابقة في هذا المجال إلا أنها لا تزال قيد البحث التطوير. إضافة إلى أن الدراسات التي تناولت اللغة العربية اقتصرت على عدد قليل من الأبحاث، إذ يبرز التحدي بشكل أكبر في اللغة العربية بسبب العدد المحدود من بيانات التدريب المنمّطة التي يمكن الاعتماد عليها لتدريب النماذج اللغوية.

3.1. فكرة الحل المقترح

نقدم في هذا البحث منهجية لبناء نظام مساعد على كشف أخطاء النطق باستخدام نوع خاص من السمات تدعى واصفات الكلام. تساعد هذه السمات في تمثيل إشارة الكلام وفق مخرج الصوت وصفاته، إذ تمثِّل هذه السمات مجتمعة كل صوت بشكل مختلف يميزه عن غيره من الأصوات، وبالتالي يمكن من خلال هذا التمثيل كشف وجود خطأ في نطق الصوت عند اختلاف أحد السمات الصوتية الممثلة له، إضافة للقدرة على تحديد نوع الخطأ بدقة مما يسهم في إعطاء المعلومات اللازمة لكيفية تصحيحه. ما يميز هذه السمات أيضاً فعاليتها في التطبيقات التي تتغير فيها الإشارات الصوتية بتغيّر المتكلم أو بوجود ضجيج في البيئة المحيطة وذلك مقارنة بالواصفات التقليدية. إلى جانب كون معظمها مشترك بين اللغات وهذا يسمح بتدريب نموذج عام للغات مختلفة. نستخدم أيضاً تقنيات التعلم العميق لتعلم هذه الواصفات وتمييز الأصوات فقد أثبتت الدراسات فعالية هذه التقنيات والقيمة العملية لها في تطبيقات تعرف الصوت [7].

4.1. مساهمات البحث

تتلخص إسهامات البحث بشكل أساسي بما يلي:

- تطبيق منهجية لكشف أخطاء النطق وفق مرحلتين يتم في المرحلة الأولى التحقق من صحة اللفظ وفي المرحلة الثانية تحديد نوع خطأ النطق في حال وجوده باستخدام نوع خاص من السمات تدعى واصفات الكلام.
- تطبيق المنهجية السابقة على مجموعة معطيات للغة العربية لم يتم تجريبها سابقاً في هذا السياق.
- تحسين نموذج تعرف واصفات الكلام باستخدام التعلم المتعدد المهام.

5.1. مخطط البحث

نستعرض في الفصل الثاني مقدمة نظرية عن موضوع البحث تضم كل من علم الأصوات وشبكات التعلم العميق، ننتقل في الفصل الثالث لتناول دراسة مرجعية للمنهجيات والأعمال التي جرى العمل عليها في إطار البحث مع المقارنة بينها. في الفصل الرابع سنقدم المقاربة المقترحة ومنهجية العمل المتبعة في كشف خطأ النطق وتحديد نوعه. وفي الفصل الخامس نعرض نتائج العمل والاختبار على عدة مجموعات للمعطيات وفق المقاربة المقترحة، ومقارنة النتائج مع الأعمال السابقة. نختم البحث في الفصل السادس بالحديث عن أهم الاستنتاجات والمقترحات لتحسين العمل مستقبلاً في الأبحاث القادمة.

نقدم في الفصل القادم الأساسيات النظرية اللازمة للبحث التي تتعلق بعلم الأصوات وواصفاته من جهة وتلك التي تتعلق بتقنيات التعلم العميق من جهة أخرى.

الفصل الثاني: الدراسة النظرية

نعرض في هذا الفصل المفاهيم الأساسية التي يقوم عليها البحث، نتحدث بشكل مفصل عن علم الأصوات ونتناول بشكل خاص الأصوات في اللغة العربية وسماتها الصوتية المرتبطة بمخارج الحروف وصفاتها. ثم نستعرض آلية معالجة إشارة الكلام وأهم السمات التي يتم استخلاصها منها. بعد ذلك ننتقل للحديث عن شبكات التعلم العميق المستخدمة لكشف السمات وتصنيف الأصوات وبعض الأساسيات فيها وأنواعها المرتبطة بالبحث مع شرح بسيط عن بنية كل منها.

1.2. علم الأصوات

1.1.2. مقدمة

تعد اللغة الوسيلة الكلامية الرئيسية للتواصل بين البشر وتناقل المشاعر والأفكار بينهم، وهي أصل الحضارة الإنسانية ورقيها وتقدمها. ولما كانت اللغة في شقها المنطوق والمكتوب نواة حياة الإنسان وتحضّره أَولى لها عناية فائقة فقام بدراستها وتفسيرها وفكّ رموزها. وهكذا برزت العديد من الدراسات اللغوية على مر العصور في حضارات مختلفة وكان أقدمها في الهند عندما قام بانيني (Panini) في القرن الخامس أو الرابع قبل الميلاد بدراسة مخارج الأصوات وتأثر بعضها ببعض في اللغة السنسكريتية [8]. كما كان للعرب دورٌ بارز في الدراسة اللغوية للغتهم حيث اهتموا بها لاهتمامهم بالقرآن الكريم وسعيهم لصون اللغة العربية من التحريف والتغيير. وفي ضوء ذلك وضع اللغويون العرب القواعد النحوية والصرفية ودرسوا جهاز النطق عند الإنسان، فقسموه إلى مخارج ونسبوا لكل مخرج مجموعة الأصوات المنتمية إليه، وصنّفوا الأصوات العربية لفئات مختلفة وفقاً لمعايير خاصة، كتقسيم الأصوات إلى صحيحة ومعتلّة ومجهورة ومهموسة ومفخّمة ومرقّقة وغير ذلك. ومن أبرز العلماء العرب القدامى الذين وضعوا أسس وقواعد اللغة: أبو الأسود الدؤلي، والخليل بن أحمد الفراهيدي، وسيبويه، وابن جُنّي وابن سينا وغيرهم... وتبعهم في منهجهم علماء اللغة المعاصرون مع ما أضافوه في دراساتهم نتيجة اتصالهم بعلماء اللغة الأجانب حيث ظهرت فروع جديدة في ميدان الدراسة اللغوية لم يتم التطرق لها سابقاً[9].

قسم اللغويون اللغة لمستويات مختلفة لتسهيل دراسة الظواهر اللغوية، منها المستوى النحوي والصرفي والدلالي والصوتي. وسنركز فيما سيأتي على المستوى الصوتي لارتباط البحث فيه.

تتم دراسة الصوت بمنهجين مختلفين [10]:

علم الأصوات اللغوية phonetics يدرس الأصوات البشرية الكلامية ويصنفها ويحلّلها ويجري عليها التجارب ويشير لكيفيّة إنتاجها وانتقالها واستقبالها، دون النظر إلى ما تنتمي إليه هذه الأصوات من لغات أو إلى وظيفة الأصوات ودورها في تغيير معنى الكلمة.

علم وظائف الأصوات اللغوية (الفونولوجيا) phonology يدرس الصّوت الإنساني من حيث وظيفته في تركيب الكلام ودوره في الدراسات الصّرفية والنحوية والدلالية في لغة معينة.

سنعتمد فيما يلي علم وظائف الأصوات اللغوية لدراسة النظام الصوتي.

2.1.2. الأصوات في اللغة العربية

تتألف الإشارة الكلاميّة من مجموعة وحدات صوتيّة تسمى **الصوتيمات phonemes**. وقد ظهر مفهوم الصوتيم لأول مرة في القرن التاسع عشر في بريطانيا وروسيا [9] وتعدّدت الدراسات والبحوث عنه، ووُضعت له العديد من التعريفات. وبشكل عام يشير مصطلح الصوتيم في علم الصوت الى أصغر وحدة صوتية في اللغة تميّز كلمة عن كلمة أخرى [10] . على سبيل المثال، يلاحظ المستمع في اللغة العربية الفرق بين كلمتي "قام" و "نام" بسبب اختلاف الوحدتين الصوتيتين (الحرفين) أو الصوتيمين القاف والنون فيهما، مع اتّفاق باقي المكونات الصوتية. وتختلف الصفات المميّزة للصوتيم من لغة الى أخرى على الرغم من اشتمال كل منها على صوتيمات متماثلة، فالباء في العربية هي صوت مجهور لا يصاحبه خروج للهواء، أما في الإنجليزية فقد يكون مجهوراً كنظيره في العربية فينطق B مثل (bay / خليج) وقد يكون مهموساً يصاحبه جريان للهواء فينطق P مثل (pay / دفع).

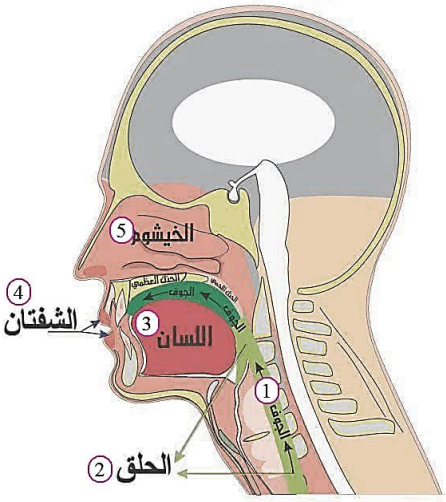
يتولد عن اجتماع الصوتيمات المقاطع الصوتية Syllables وعلى مستوى أعلى الكلمات والجمل. سنركز في دراستنا فيما سيأتي على الأصوات العربية كونها محور البحث، وسنستخدم ترميز المنظمة العالمية للصوتيات (International Phonetic Association) لتوضيح رموز الأصوات. تسمى هذه الرموز الأبجدية الصوتية العالمية (International Phonetic Alphabet) أو اختصاراً IPA.

تقسم الأصوات اللغوية فونولوجياً الى قسمين رئيسيين: الصوامت (consonants) والصوائت (vowels). فالصوامت هي /ء/، /ب/، /ت/، /ث/ ...، أما الصوائت فهي الحركات وما يقابلها من حروف المد (الفتحة والألف المدية، الضمة والواو المدية، الكسرة والياء المدية)، ولا تُعد الواو والياء من الصوائت إلا عندما تُسبَق بحركة من نفس جنس الحرف (الياء قبلها كسر والواو قبلها ضم). تصدر الأصوات نتيجة حركة الهواء عبر الفم أو الحلق أو الأنف، والفرق بين الصوامت والصوائت أن الصوامت يصاحبها تقارب بين عضوي النطق مما يؤدي لمنع تدفق الهواء الخارج من الرئتين فيها بشكل جزئي مثل /ث/ /ف/ /ع/،

أو بشكل كلي /ب/ /ت/ /ق/. أما الصوائت فتكون درجة اقتراب عضوي النطق فيها أقل ويكون التجويف الفموي أثناء نطقها مفتوحاً مثل حروف المد وتتميز بطول زمنها مقارنة بغيرها.

تحوي اللغة العربية أربعة وثلاثين صوتاً منها ستة صوائت (حروف المد والحركات المقابلة لها)، و28 صامتاً، وتصنف هذه الأصوات حسب مخرج النطق Place of Articulation (الجوف، الحلق، اللسان، الشفتان، الخيشوم) (انظر الشكل 1)[1] أي حسب المكان من الجهاز الصوتي الذي يخرج منه صوت الحرف [9] ,[8]، إلى:

1. الأصوات الشفوية labial هي الأصوات التي تخرج من الشفتين، ويمكن تفصيل هذه الأصوات إلى:
   - شفوي ثنائي bilabial تخرج من بين الشفتين وهما /م/ و /ب/
   - شفوي أسناني labiodental تخرج من بين الثنايا العليا والشفة السفلى وهو /ف/
   - شفوي طبقي labiovelar تخرج بانضمام الشفتين مع ارتفاع مؤخر اللسان للأعلى /الواو غير المدية/



الشكل 1– مخارج الحروف الرئيسية

2. الأصوات الحلقية pharyngeal هي الأصوات التي تخرج من الحلق برجوع لسان المزمار إلى جدار الحلق وهي /ع/ /ح/
3. الأصوات اللهوية uvular تخرج من أقصى اللسان من المنطقة الواقعة بين اللهاة ومؤخر اللسان /ق/ /غ/ /خ/
4. الأصوات الطبقية velar تخرج من أقصى اللسان من المنطقة الواقعة بين الحنك ومؤخر اللسان /ك/
5. الأصوات الحنجرية glottal تخرج من منطقة الأوتار الصوتية وهي /هـ/ /ء/

[1] https://www.diveintoarabic.com/en/how-to-say-in-arabic-en/letters-articulation-points/

6. الأصوات الغارية palatal تخرج من بين وسط اللسان وغار الحنك الأعلى /ج/ /ش/ /الياء غير المدية/ ويسمى مخرج كل من /ج/ /ش/ حنكي لثوي alveolo-palatal .
7. الأصوات بين الأسنانية interdental تخرج من بين أطراف الثنايا العليا وطرف اللسان وهي /ظ/ /ذ/ /ث/
8. الأصوات اللثوية alveolar تخرج باشتراك اللثة مع عضو نطق آخر وتصنف إلى:
   - حنكي لثوي alveolo-palatal /ج/ /ش/ كما مر معنا سابقاً.
   - لثوي أسناني alveo-dental تخرج من طرف اللسان مع ما يقابله من اللثة /ت/ /ط/ /د/ /ص/ /س/ /ز/ /ل/ /ن/ /ر/ أو من حافة اللسان مع الأضراس العليا /ض/، وتصنف إلى:
   - ❖ الأصوات النطعية pre-palatal تخرج من طرف اللسان (النطع) مع ما يقابله من لثة الثنايا العليا في منطقة جذور الأسنان /ط/ /د/ /ت/
   - ❖ الأصوات الأسلية laminal تخرج من طرف اللسان (الأسلة) مع ما يقابله من الثنايا السفلى من الداخل /ص/ /س/ /ز/
   - ❖ الأصوات الذلقية liquids تخرج من ذلق اللسان(طرفه) /ل/ /ن/ /ر/

9. الأصوات الجوفية cavity تخرج من الجوف وهي حروف المد /ا/ /و/ /ي/ حركتها السكون وما قبلها موافق لها، فالواو ساكنة قبلها ضم، والياء ساكنة قبلها كسر، والألف ساكنة قبلها فتح، ويلحق بهذه الحروف الحركات التابعة لها (الفتحة والضمة والكسرة).
10. الأصوات الأمامية anterior تخرج من مقدم الفم. /س/ /ف/ /ت/ /ن/ /ز/ /ذ/ /و/ /ظ/ /ث/ /ل/ /ر/ /ض/ /ط/ /م/ /د/ /ب/ /ص/.
11. الأصوات الوسطية coronal تخرج من طرف اللسان ومقدمه ووسطه وما يقابله من الثنايا والحنك. /س/ /ت/ /ن/ /ذ/ /ز/ /ش/ /ظ/ /ث/ /ل/ /ر/ /ض/ /ط/ /د/ /ص/ /الفتحة/ /ا/.
12. الأصوات المدورة rounded يصاحب نطقها تدوير للشفتين /و/ /الضمة/.

إن مخارج الحروف غير كافية للتمييز بين الأصوات فبعض الأصوات لها المخرج نفسه (كالمخرج اللثو_أسناني والذي يخرج منه 10 أصوات مختلفة)، والفارق بينها وجود صفات مميّزة لها تتحدد بآلية نطق هذه الحروف تسمى كيفية النطق Manner of Articulation ويعبر عنها بصفات الحروف [8]، وتصنف كما يلي:

1. الأصوات المجهورة voiced: يهتز فيها الوتران الصوتيان نتيجة اقترابهما من بعضهما البعض وضيق مجرى الهواء عند النطق بالصوت. وتندرج الصوائت إضافة للصوامت التالية ضمن الأصوات المجهورة /ب/ /ء/ /ج/ /د/ /ذ/ /ر/ /ز/ /ض/ /ط/ /ظ/ /ع/ /غ/ /ق/ /ل/ /م/ /ن/ /و/ /ي/.

تقابل الأصوات المجهورة الأصوات المهموسة voiceless والتي لا يهتز فيها الوتران الصوتيان نتيجة تباعدهما واتساع مجرى الهواء بينهما، وهي الصوامت التالية /ف/ /ح/ /ث/ /هـ/ /ش/ /خ/ /ص/ /س/ /ك/ /ت/. أي أنها جميع الأصوات العربية عدا المجهورة منها.

2. الأصوات الانفجارية (stops (plosives ينغلق فيها مجرى الهواء تماماً لفترة قصيرة يتبعها صوت انفجاري عند انفتاح المخرج ومرور الهواء. وهي نوعان:
   - أنفية nasal stops: ينغلق فيها تجويف الفم ليمر الصوت عبر التجويف الأنفي وتكون في حرفي /م/ /ن/.
   - فموية oral stops: ينحبس فيها جريان الصوت والنفس عبر التجويفين الفموي والأنفي بسبب انغلاق المخرج بالكامل إما باللسان (ك ت..) أو الشفتين (ب). تسمى هذه الأصوات بالأصوات الشديدة fortis، وتتصف بصفة الشدة strength وهي /ق/ /ط/ /ب/ /ج/ /د/ /ك/ /ت/ /ء/ تجمعها عبارة /أجد قط بكت/، تأتي بدرجة أقل منها صفة التوسط moderate والتي يجري فيها الصوت بشكل جزئي بسبب عدم كمال غلقه، وهي الأصوات المجتمعة في عبارة /لن عمر/، أما الأصوات الرخوة lenis فتتصف بالرخاوة softness، يجري فيها الصوت بشكل تام عند مروره في المخرج وتشمل بقية الأصوات.

3. الأصوات الاحتكاكية fricatives ينحبس فيها النفس بشكل جزئي فينتج صوت خفيّ وتسمى هذه الأصوات بالرخوة وهي /ف/ /ث/ /ظ/ /س/ /ص/ /ز/ /ش/ /خ/ /ح/ /ع/ /هـ/ /ذ/.
4. الأصوات المركبة affricates هي صوت وقفي شديد متبوع بصوت احتكاكي وهو في اللغة العربية الصوت المعاصر الفصيح للجيم.
5. الأصوات المفخمة emphatic يصاحب هذه الأصوات ارتفاع أقصى اللسان للحنك الأعلى ويتميّز الصوت بتصعده لأعلى الحنك بحيث ينحصر الصوت بين اللسان والحنك، وتتصف هذه الأصوات بصفة الإطباق adhesion وهي /ص/ /ض/ /ظ/ /ط/ ، أما باقي الأصوات فتسمى بالأصوات المنفتحة (لها صفة الانفتاح separation) أي لا يطبق معها اللسان على الحنك الأعلى، بل يبقى في وضع الراحة دون ارتفاع.
6. الأصوات التكرارية trills يتكرر فيها اتصال عضو نطق بآخر حيث ينتج عن كل عملية اتصال حرف وهذه الصفة خاصة بحرف واحد هو الراء يلامس فيه طرف اللسان اللثة العليا لوقت قصير لينتج صوت الحرف ويؤدي تكرار العملية لتكرار صوت الحرف repetition.
7. الأصوات الانتشارية spread ينتشر فيها الهواء في الفم عند النطق بها، وهي تخص صوت /ش/ وتتصف هذه الأصوات بصفة التفشي.

8. الأصوات الجانبية lateral يصاحبها وقف لجريان الهواء في الفم وهي نوعان:

- جانبية تقاربية lateral approximants يبتعد فيها جانبا اللسان عن الحنك مما يسمح للصوت بالمرور عبرهما، تخص صوت /ل/ ويعد من الأصوات المنحرفة[2] التي ينحرف فيها الصوت عن مخرجه لاعتراض اللسان طريقه، إضافة لصوت الراء.
- جانبية احتكاكية lateral fricatives يقترب فيها أحد جانبي اللسان أو كلاهما من الحنك، مما يؤدي إلى انحباس النفس واضطراب الهواء في الفم، وتتصف هذه الأصوات بالاستطالة prolongation أي أن الصوت يمتد فيها من أول حافة اللسان الى آخرها، ويتميز /ض/ بهذه الصفة.

9. الأصوات التقاربية approximant يتقارب فيها عضوا النطق ولا ينغلقان بشكل كامل مما يسمح للهواء بالمرور أكثر من مروره في الأصوات الاحتكاكية وهي نوعان:

- تقاربية جانبية lateral approximants هي التي ذكرت سابقاً وتكون في /ل/.
- تقاربية وسطية central approximants تكون في /و/ /ي/ غير المدية حيث يقترب وسط اللسان في الياء من الحنك الأعلى. بينما في الواو يقترب أقصى اللسان من الحنك الليّن إضافة للتقارب بين الشفتين.

10. الأصوات الصفيرية sibilant هي أصوات احتكاكية لها نبرة عالية تنتج عن مرور الهواء في منفذ ضيق يميزها عن الأصوات الاحتكاكية الأخرى، وهي /ص/ /س/ /ز/ وتتصف هذه الأصوات بصفة الصفير whistle.

11. الأصوات المستعلية elevated (high) تشمل الأصوات المفخمة إضافة للغين والخاء والقاف حيث يصحبها ارتفاع أقصى اللسان للحنك الأعلى وتتصف بالاستعلاء elevation. تقابلها صفة الاستفال lowering مع باقي الأصوات ولا يكون فيها أي تفخيم في صوت الحرف.

12. الأصوات الخفية hiding هي الأصوات التي يخفى فيها الصوت عند النطق به، وتضم الهاء وحروف المد الثلاثة الألف والواو والياء.

---

[2] الأصوات المنحرفة deviated ينحرف فيها الصوت عن مخرجه لاعتراض اللسان طريقه وهي /ل/ و /ر/

يبين الجدول 1 تصنيف الصوامت في اللغة العربية الفصحى حسب المخارج والصفات مع ما يقابلها من رموز IPA [8]. الأصوات المجهورة تقع على يمين العمود بينما تقع المهموسة على يساره.

الجدول 1 أصوات اللغة العربية الفصحى

	Bilabial شفتاني	Labiodental شفوي أسناني	Interental بين أسناني	Alveodental لثوي أسناني	Alveopalatal غاري لثوي	Palatal غاري	Velar طبقي	Lab-velar شفوي طبقي	Uvular لهوي	Pharyngeal حلقي	Glottal حنجري
Nasal أنفي	م m			ن n							
Stop شديد	ب b			د d ت t			ك k		ق q		ء ʔ
Emphatic Stop*				ضˤ dˤ طˤ tˤ							
Fricative رخو		ف f	ذ ð ث θ	ز z س s	ش ʃ				غ ʁ خ χ	ع ʕ ح ħ	هـ h
Emphatic fricative**			ظˤ ðˤ	صˤ sˤ							
Affricate مزجي					ج dʒ						
Glide ليني						ي j		و w			
Lateral جانبي				ل l							
Trill تكراري				ر r							

ويبين الجدول 2 و 3 تصنيف لأصوات اللغة العربية حسب مخارج الحروف وصفاتها وفق مرجع آخر [11].

الجدول 2 مخارج الحروف لأصوات اللغة العربية

مخارج الحروف	الحروف
بلعومي Pharyngeal	ع ح
حلقي (بلعومي) Pharynx	ء هـ ع ح غ خ
حنجرية Glottal	ء هـ
لهوية Uvular	ق
فوق اللثة Post-alveolar	ج ش
وسط اللسان Middle tongue	ج ش ي
اقصى اللسان Deep tongue	ك ق
حافة اللسان Tongue border	ض ل
طرف اللسان Tongue tip	ط د ت ص س ز ظ ذ ث ن ر
بين الأسنان Interdental	ظ ث ذ

الجدول 3 صفات الحروف لأصوات اللغة العربية

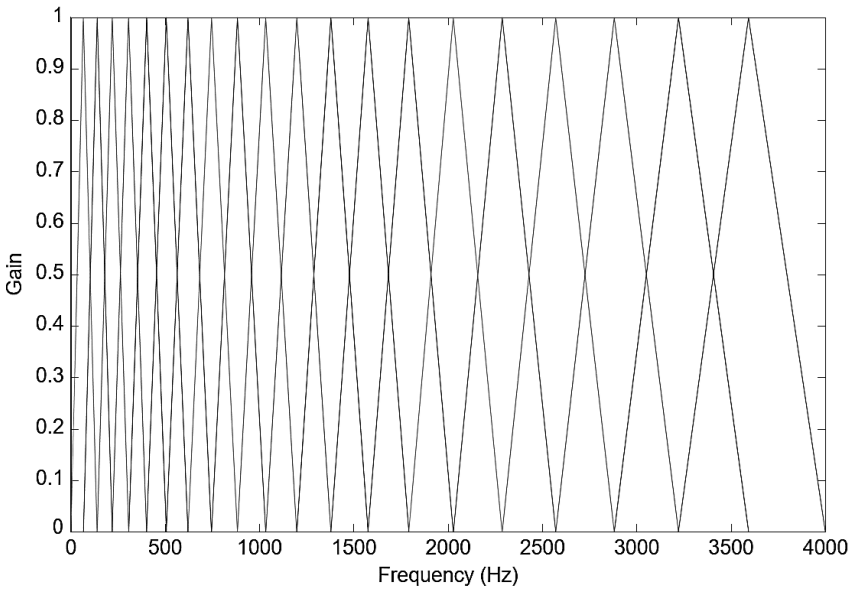
صفات الحروف	الحروف
الهمس Whisper	فحثه شخص سكت
الصفير Whistle	ص س ز
الإطباق Adhesion	ص ض ط ظ
مركبة Affricates	ج
الانحراف Deviate	ل ر
الاستعلاء Elevation	خص ضغط قظ
احتكاكية Fricatives	ص س ز ظ ذ ه ح ع غ خ ش ف ث
الرخاوة Softness	و ي ض ف غ خ ح ه ص س ش ذ ث ظ ز (ا و ي) المدية
التفشي Spreading	ش
الشدة Strength	أجد قط بكت
الخفاء Hiding	هاوي
التوسط Moderate	لن عمر
التكرار Repetition	ر
الاستطالة Prolongation	ض
الصوائت Vowels	(ا و ي) المدية
الصمت Silence	–

3.1.2. معالجة الإشارة الصوتية

تتطلب جميع مجالات تطبيق تقانات الكلام، بما في ذلك تعرف الكلام وتركيبه وترميزه ، شكلاً من أشكال التحليل الأولي لإشارة الكلام [12]. ويكون الهدف الرئيسي في عملية التحليل تقدير الاستجابة الترددية للقناة الصوتية. يعد مفهوم "تحليل على المدى القصير short-time " أساسياً لمعظم تقنيات تحليل الكلام. حيث يفترض أن شكل موجة الكلام يكون غير ثابت على مدى فترة زمنية طويلة، لكن يمكن اعتباره مستقراً خلال فترة زمنية قصيرة (10–30) مللي ثانية هي وسطياً زمن الصوتيم. ويرجع ذلك إلى أن المعدل الذي يتغير فيه طيف الكلام يرتبط ارتباطًا مباشرًا بمعدل حركة مفاصل الكلام (الشفاه واللسان والفك وما إلى ذلك). وبالتالي، فإن معظم أنظمة تحليل الكلام تعمل على أساس متغير زمنياً، باستخدام مقاطع قصيرة من الكلام مختارة على فترات زمنية متباعدة بشكل موحد أو إطارات ذات مدة نموذجية (10–30) مللي ثانية. للحصول على إشارة الكلام الرقمية، يمكن أخذ العينات بمعدل 16 كيلو هرتز. ومن الناحية العملية، يكفي معدل أخذ العينات 16 كيلو هرتز لعرض طيف الكلام (8 كيلو هرتز).

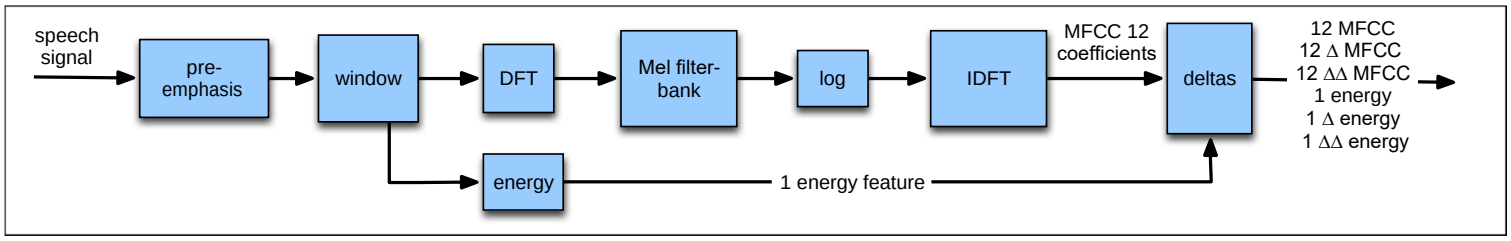
إن معدل أخذ العينات المنخفض، يزيد بشكل عام من معدل أخطاء تعرف الكلام. كما أن زيادة معدل أخذ العينات بشكل كبير ليس له أي تأثير إضافي على أخطاء تعرف الكلمات، لأن معظم ميزات الكلام البارزة تقع ضمن عرض النطاق الترددي 8 كيلو هرتز[10].

بعد تقسيم الإشارة لإطارات، يتم استخلاص مجموعة من المعاملات أو السمات لكل إطار للحصول على معلومات حول جهاز النطق والوظيفة الخاصة به بشكل مباشر أو غير مباشر، إضافة إلى معلومات تميز شكل موجة الكلام نفسها. يعد استخلاص السمات (Feature Extraction) من أهم مواضيع تعرف الكلام، ويتم فيه تحويل موجة الدخل إلى تسلسل من أشعة السمات الصوتية، يمثل كل شعاع معلومة عن نافذة زمنية قصيرة من الإشارة [13]. يوجد عدد كبير من السمات التي يمكن استخدامها، وتعد معاملات ميل التردّدية (Mel-frequency cepstral coefficient) MFCC أشهرها كونها تُنمذج إدراك الإنسان للصوت، وتستند إلى تحليل التردد الذي يتم إجراؤه في الأذن الداخلية [14]. الفكرة الأساسية في حساب MFCC هي الحصول على المعلومات الترددية بالاعتماد على مجموعة مرشحات filter bank لمعرفة كمية الطاقة المختزنة في الإشارة عند نطاقات ترددية مختلفة عن طريق إجراء تحويل فورييه المتقطع discrete Fourier transform (DFT). يتم استخدام 20 مرشح تقريباً لعرض نطاق ترددي 4 كيلو هرتز. لحساب DFT تستخدم خوارزمية تحويل فورييه السريع المعروفة Fast Fourier transform (FFT)، بعد ذلك يتم تجميع قيم DFT معاً في نطاقات دقيقة بمعدل 10 مرشحات موزعة بشكل خطي في المجال 0−1000 هرتز، وبقية المرشحات توزع بشكل لوغاريتمي في المجال الأعلى من 1000 هرتز، ليتم تطبيق توابع ترجيح مثلثية triangular functions عليها (الشكل 2) ، في المرحلة التالية يتم حساب لوغاريتم القيم الناتجة، وهذا يساعد بمحاكاة الاستجابة السمعية ويجعل عملية تقدير السمات أقل حساسية لتغيرات الدخل (كتغير الطاقة الناتج عن اقتراب فم المتحدث أو بعده عن الميكروفون خلال الكلام). يتم بعد ذلك حساب معاملات كبستروم Cepstrum عن طريق تحويل فورييه المتقطع العكسي inverse Discrete Fourier Transform (IDFT) بهدف الحصول على السمات التي تساعد في التمييز بين الصوتيمات المختلفة، وهكذا نكون حصلنا على معاملات MFCC للإشارة الصوتية.



الشكل 2– مجموعة مرشحات مثلثية Triangular filterbank

إن الإشارة الصوتية لا تكون ثابتة من إطار لآخر بسبب طبيعة التغيرات التي تحدث عند نطق الصوت، لهذا يتم استخدام سمات إضافية قادرة على ضبط التغير بين الإطارات الزمنية وهي السرعة (دلتا delta) والتسارع (دلتا دلتا delta delta). وهما المشتق الزمني الأول والثاني لمعاملات MFCC على التتالي. إضافة للسمات السابقة تُستخلص الطاقة كونها ترتبط بهوية الصوتيم، فهي إشارة مفيدة لتمييزه. على سبيل المثال تمتلك حروف المد وحروف الصفير طاقة أعلى من الحروف الوقفية. وتحسب الطاقة في كل إطار صوتي كمجموع مربعات العينات على نافذة التحليل. يوضح (الشكل 3) خطوات استخلاص السمات الصوتية السابقة بفرض تم أخذ 12 قيمة طيفية، فيكون شعاع السمات الناتج بطول 39 سمة. بالنسبة للسمات المستخدمة في بحثنا سنذكرها بالتفصيل في فصل المقاربة المقترحة.



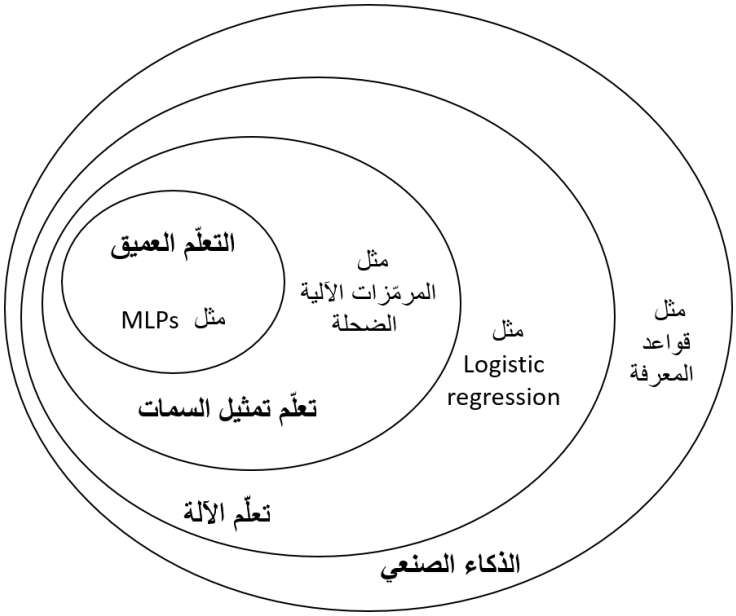
الشكل 3– استخلاص شعاع سمات MFCC من الإشارة الصوتية

2.2. شبكات التعلم العميق

ظهر مفهوم الذكاء الصنعي في منتصف القرن العشرين مع سعي الباحثين للحصول على أنظمة ذكية تقوم بأتمتة الأعمال الروتينية وجعل الحواسيب قادرة على القيام بالمهام التي يقوم بها الإنسان بشكل بديهي كفهم الكلام والتعرف على الأشياء والأشخاص. عملت أنظمة الذكاء الصنعي في بداياتها ضمن بيئات بسيطة عن طريق قواعد معرفة مسبقاً لا تحتاج من الحواسيب معرفة كبيرة عن العالم الخارجي، ومع تطور هذه الأنظمة ظهرت الحاجة لإيجاد طريقة يقوم الحاسوب من خلالها باستخلاص الأنماط من البيانات الخام المدخلة له، وبالتالي يصبح قادراً على القيام بمهام معقدة أكثر واتخاذ القرارات المناسبة بنفسه بطريقة ذكية وهو ما يسمى **بتعلّم الآلة machine learning**. تم تطوير العديد من خوارزميات تعلّم الآلة لحل مسائل مختلفة، واعتمد أداء هذه الخوارزميات على الطريقة التي يتم بها تمثيل المسألة بمجموعة من السمات features حيث تختلف هذه السمات باختلاف المسألة المطروحة مما شكل صعوبة لاحقاً لصعوبة معرفة السمات المناسبة التي يجب استخلاصها من البيانات لمختلف المهام. لحل هذه الصعوبات تم اقتراح استخدام تعلّم الآلة لاكتشاف التمثيل المناسب للسمات representation learning بدلاً من الاعتماد على التمثيلات المصممة يدوياً والتي تحتاج لجهد كبير إضافة للكثير من الوقت للحصول عليها، وهذا يسمح لأنظمة الذكاء الصنعي بالتكيف مع مسائل جديدة بسهولة.

إن تصميم خوارزمية تعلم لتمثيل السمات يأخذ بالاعتبار وجود عوامل للتباين في البيانات المدروسة، على سبيل المثال في مهمة التعرف على سيارة في الصورة تحدد عوامل التباين بموقع السيارة ولونها وحجمها وزاويتها، وهذا يتطلب تمثيل البيانات بشكل مجرد لتحقيق المهمة المطلوبة بحيث يتم الانطلاق من تمثيل بسيط للمفاهيم لبناء مفاهيم أكثر تعقيداً، وهذا ما يسمى **بالتعلّم العميق Deep learning**، فهو طريقة لتمثيل البيانات بمستويات عديدة عن طريق تكوين وحدات بسيطة غير خطية (طبقات) يقوم كل منها بتحويل التمثيل على مستوى واحد وصولاً لمستويات تمثيل أعلى بتجريد أكبر وذلك انطلاقاً من الدخل الخام للبيانات [15]. فمثلاً يكون دخل نظام التعلم في الصور مصفوفة تحوي قيم البكسلات، وتحوي طبقة التمثيل الأولى عادة الواصفات التي تعلمها النظام والتي تمثل وجود أو غياب الحواف في مواقع معينة وبزوايا معينة من الصورة، بينما قد تحدد الطبقة الثانية أشكال وأنماط مختلفة وفق ترتيبات معينة للحواف، بغض النظر عن الاختلافات الصغيرة في مواضع الحواف. في الطبقة الثالثة من الممكن أن يتم تجميع الأنماط من الطبقة السابقة في مجموعات أكبر تتوافق مع أجزاء من الأشياء المألوفة، والطبقات اللاحقة ستكتشف الأجسام في الصورة كمجموعات من هذه الأجزاء. في التعلم العميق لا تُصمَّم طبقات السمات من قبل البشر وإنما يتم تعلمها من البيانات باستخدام إجراءات تعلم عامة الغرض. ولهذا حقق التعلم العميق تقدماً كبيراً في حل كثير من المشاكل التي عجز عنها مجتمع الذكاء الصنعي لسنوات عديدة.

يوضح (الشكل 4) العلاقة بين تخصصات الذكاء الصنعي المختلفة [16] ، نجد أن التعلم العميق هو جزء من تعلم التمثيل، والذي يعد أيضاً جزء من تعلّم الآلة، ويُستخدم في العديد من تطبيقات الذكاء الصنعي



الشكل 4– مخطط فين للعلاقة بين تخصصات الذكاء الصنعي المختلفة

1.2.2. تعريف شبكة التعلم العميق

هي شبكة عصبونية بعدة طبقات خفية، تتميز عن الشبكات العصبونية التقليدية بكونها أكثر تعقيداً وبقدرتها على توفير حلول للعديد من مشكلات التعلم، إضافة لقدرتها على معالجة كميات ضخمة من البيانات بأبعاد عالية. تتكون الشبكة العصبونية من طبقة دخل (Input layer)، وطبقة خرج (output layer)، ومجموعة من الطبقات الخفية (Hidden layers). تؤلف سمات الدخل طبقة الدخل، بينما تمثل طبقة الخرج الطبقة الأخيرة في النموذج، ويُحدَّد عدد العصبونات المرتبطة بهذه الطبقة بعدد المخارج التي نريد الحصول عليها. بالنسبة للطبقات الخفية التي تربط طبقة الدخل بالخرج فهي لَبِنة أساسية في بناء الشبكة حيث تسمح بنمذجة البيانات المعقّدة لوجود العصبونات (العقد nodes/neurons) فيها، وتسمى خفيّة لأن القيمة الحقيقية للعقد في بيانات التدريب تكون غير معروفة، إذ تقتصر معرفتنا على كل من الدخل والخرج فقط. تعد العصبونات وحدة المعالجة في الشبكة، يجمع العصبون بين مدخلات البيانات ومجموعة من المعاملات أو الأوزان التي تضخم أو تثبط المدخلات، وبالتالي تعطي أهمية للمدخلات حسب المهمة التي تحاول الخوارزمية تعلمها. يوجد للعصبون تابع تفعيل (activation function) يحدد ما إذا كان يجب تمرير الإشارة عبر الشبكة والمقدار الذي يتم تمريره للتأثير على النتيجة النهائية.

2.2.2. أساسيات في التعلم العميق

تتكون شبكة التعلم بشكل أساسي، وبغض النظر عن نوع مشكلة تعلّم الآلة التي نتعامل معها من المكونات التالية [17]:

- البيانات التي يمكننا التعلم منها.
- نموذج البيانات.
- تابع الهدف الذي يحدد مدى جودة (أو سوء) النموذج.
- خوارزمية ضبط موسطات النموذج لتحسين تابع الهدف.

❖ **البيانات Data**

تعد البيانات اللبنة الأساسية في عملية التعلّم فلا يمكن أن يقوم علم البيانات دونها. وحتى نعمل مع البيانات بشكل مفيد نحتاج لتمثيلها بشكل رقمي مناسب، بحث تتكون كل عينة بيانات من مجموعة ميّزات تسمى السمات features (أو المتغيرات المشتركة covariates) يقوم النموذج بالتنبؤ بها. عندما يتم تمثيل كل عينة بنفس عدد القيم، نقول إن البيانات تتكون من متجهات ذات طول ثابت ونَصِفُ الطولَ الثابت للمتجهات بأبعاد البيانات، لكن لا يمكن بسهولة تمثيل جميع البيانات كمتجهات ثابتة الطول، فمثلاً الصور لها أبعاد مختلفة تتغير بتغير الدقة والشكل. أيضاً البيانات النصية والصوتية لها أطوال متغيرة.

تتمثل إحدى الميزات الرئيسية للتعلم العميق عن الطرق التقليدية في قدرتها على التعامل مع بيانات متفاوتة الطول.

❖ **النماذج Models**

يتضمن تعلم الآلة تحويل البيانات بطريقة معينة للقيام بمهمة محددة. كبناء نظام يقوم بمعالجة الصور ويتعرف على الأشخاص. يحدد النموذج الآليات الحسابية التي ستتم للقيام بمعالجة مجموعة البيانات والحصول على التنبؤات المناسبة، ونركز في تعلم الآلة على النماذج الإحصائية التي يمكن تقديرها من البيانات. يختلف التعلم العميق عن الأساليب التقليدية بقوة النماذج التي يتم بناؤها. تتكون هذه النماذج من عدة تحويلات متتالية للبيانات يتم ربطها ببعضها البعض، وهذا سبب إطلاق اسم التعلم العميق عليها.

❖ **توابع الهدف Objective Functions**

إن تطوير نظام رياضي لتعلّم الآلة يحتاج إلى وجود مقياس يعبر عن مدى جودة (أو سوء) النموذج وهذا ما يدعى تابع الهدف. اصطلاحاً يحدد تابع الهدف بحيث يكون التابع ذو القيمة الأقل هو الأفضل، ويسمى بتابع الخسارة loss function. عند التنبؤ بالقيم العددية، فإن تابع الخسارة الأكثر شيوعًا هو الخطأ التربيعي (مربع الفرق بين قيمة التنبؤ والقيمة الحقيقة). أما بالنسبة للتصنيف، يكون الهدف الأكثر شيوعًا تقليل معدل الخطأ (نسبة الأمثلة التي لا توافق فيها قيمة التنبؤ القيمة الحقيقة).

يتم تحديد تابع الخسارة عادةً وفق موسطات النموذج وبالاعتماد على مجموعة البيانات. ويجري تعلم أفضل قيم لموسطات النموذج من خلال تقليل الخسارة الحاصلة على مجموعة أمثلة التدريب. ومع ذلك، فإن الأداء الجيد مع بيانات التدريب لا يضمن أننا سنعمل بشكل جيد على البيانات غير المرئية للنموذج. لذلك يتم تقسيم البيانات عادة إلى قسمين: مجموعة بيانات التدريب (أو مجموعة التدريب، لضبط موسطات النموذج)، ومجموعة بيانات الاختبار (أو مجموعة الاختبار، التي يتم الاحتفاظ بها للتقييم)، ليتم اختبار أداء النموذج على كل منهما.

❖ **خوارزميات التحسين (الأمثَلَة) Optimization Algorithms**

بعد الحصول على البيانات وتمثيلها، وتحديد النموذج، وتابع الهدف، نحتاج إلى خوارزمية قادرة على البحث عن أفضل الموسطات الممكنة لتقليل تابع الخسارة. تعتمد خوارزميات التحسين الشائعة للتعلم العميق على منهجية تدعى التدرج المنحدر gradient descent. تتحقق هذه الطريقة في كل خطوة لكل موسط، من مقدار تغير خسارة مجموعة التدريب إذا قمنا بتعديل قيمة الموسط بمقدار صغير جداً. ليتم تحديث قيمة الموسط بشكل يقلل من الخسارة.

3.2.2. اعتبارات عملية

يتطلب تعليم نموذج بكفاءة وفعالية أخذ العديد من القضايا العملية [18] سنتطرق لذكر بعضها فيما يلي:

❖ **المعالجة الأولية للبيانات Data Preprocessing**

تلعب عملية معالجة البيانات دوراً هاماً في العديد من خوارزميات تعلّم الآلة، والتقنيتان الأكثر شيوعاً في عمليات المعالجة هما: استنظام السمات لكل عينة (per-sample feature normalization)، وتقييس السمات للعينات بشكل عام (global feature standardization). بالنسبة للاستنظام يتم طرح متوسط السمة من السمة عندما يعطي المتوسط انحرافاً(تغيراً) لا يرتبط بالمسألة التي يتم العمل عليها، وذلك لتقليل تباين السمات التي تدخل لشبكة التدريب. فمثلاً طرح متوسط كثافة الصورة يمكن أن ينقص التباين الناتج عن الإضاءة، وفي تطبيقات التعرف على الكلام تستخدم تقنية استنظام متوسط الطيف CMN (Cepstral Mean Normalization) لطرح متوسط سمات MFCC لكل صوت من أُطر الصوت مما يخفض تشوهات القناة الصوتية. أما بالنسبة للتقييس العام للسمات فيكون بهدف ضبط مجال قيم البيانات باستخدام تحويلات عامة بحيث تصبح قيم البيانات في نفس المجال الرقمي. على سبيل المثال يتم تقييس السمات ليكون لها متوسط صفري وتباين موحد في تطبيقات الصوت.

❖ **تهيئة النموذج Model Initialization**

تبدأ خوارزمية التعلّم من نموذج ابتدائي. يوجد العديد من الطرق التجريبية التي يمكن من خلالها تهيئة نموذج شبكة التعلم وتستند هذه الطرق لاعتبارين أساسيين: الأول هو الأوزان التي يجب تهيئتها ليعمل

كل عصبون في المجال الخطي لتابع sigmoid (انظر توابع التفعيل لاحقاً) وذلك في بداية عملية التعلم. فإذا كانت قيم الأوزان عالية جداً ستكون قيم العديد من العصبونات قريبة من 0 أو 1 وبالتالي تكون قيم المشتقات صغيرة جداً، أما عندما تعمل العصبونات في المجال الخطي فتصبح قيم المشتقات عالية كفاية لتجري عملية التعلم بشكل فعال أكثر. الاعتبار الثاني هو عملية تهيئة الموسطات بشكل عشوائي وذلك لأن العصبونات في شبكة التعلم متماثلة، فإذا كان لموسطات النموذج نفس القيم، سيكون خرج عصبونات الطبقة الخفية نفسه، وبالتالي سيتم الكشف عن نفس أنماط السمات في الطبقات الأولى. لذلك تساعد التهيئة العشوائية على التخلص من مشكلة التماثل.

❖ **اختيار قياس الدفعة Batch Size Selection**

تمثل الدفعة عدد عينات التدريب التي يتم إجراء الحسابات عليها. يؤثر قياس الدفعة على سرعة التقارب في الوصول للنتيجة والنموذج الناتج. أبسط الطرق إدخال كامل مجموعة التدريب كدفعة واحدة. تسمى طريقة استخدام الدفعات بالتدريب الدفعي batch training ولها عدة ميّزات، أولها أنها تحسّن خاصية التقارب، ثانياً أن العديد من تقنيات التسريع (مثل conjugate gradient[19] وL-BFGS[20]) تعمل أفضل مع استخدام التدريب الدفعي. وأخيراً فإن التدريب الدفعي يمكن أن يتم على التوازي عبر عدة أجهزة حاسوبية. تتطلب هذه الطريقة إتمام مراحل التدريب على كامل البيانات قبل تحديث موسطات النموذج، وهذا يجعل العملية غير فعالة في المسائل الواسعة النطاق. لذا تستخدم تقنية أخرى تدعى التدرج المنحدر العشوائي stochastic gradient descent (SGD)[21] ويطلق عليها في الأدبيات بالتعلم الآني Online learning، يتم في هذه الطريقة تحديث موسطات النموذج بالاعتماد على التدرجات المحسوبة من عينة تدريب واحدة، وهذه ميزة هامة مقارنة بطريقة التدريب الدفعي، كون التابع الهدف له عدة نقاط مثلى محلية قد يكون معظمها غير جيد، لذا فإن التدريب الدفعي سيجد الحل الأمثل حسب موسطات النموذج الأولية ويكون النموذج الناتج معتمداً بشدة على عملية التأهيل. بينما تتخلص خوارزمية SGD من هذه المشكلة بالخروج من النقاط المحلية غير الجيدة وبالتالي تمكن من تحسين النموذج خلال عملية التدريب. إن خوارزمية SGD أسرع من التدريب الدفعي وخاصة مع مجموعات المعطيات الكبيرة، لكن من الصعب جعلها تعمل على التوازي. أيضاً لا يمكن لهذه الخوارزمية أن تتقارب بشكل كامل نحو القيمة المثلى وإنما تتذبذب حولها وهذا السلوك غير مرغوب به في معظم الحالات. يتم الجمع بين الخوارزميتين السابقتين بخوارزمية تدريب الدفعات الصغيرة minibatch training التي تقدّر التدرّج على دفعة صغيرة مختارة عشوائياً من بيانات التدريب. تسمح هذه الطريقة بالتدريب على التوازي ضمن الدفعة الصغيرة وهذا يجعل التقارب أسرع.

❖ **بنية الشبكة**

تعد بنية الشبكات أحد الموسطات الهامة التي تحتاج للتحديد عند بناء شبكة التعلم. تعد كل طبقة في الشبكة كمستخلص سمات للطبقة التي تسبقها، لذا يجب أن يكون عدد العصبونات في كل طبقة كبير بشكل كافٍ لالتقاط الأنماط الأساسية، خاصة بالنسبة للطبقات الأولية كون سماتها أكثر تنوعاً وتتطلب عدداً كافياً من العصبونات لنمذجة الأنماط مقارنة بالطبقات الأخرى. بالرغم من ذلك، إذا كان حجم الطبقة كبيرًا جدًا، يصبح من السهل حدوث فرط التلبيق[3] overfit مع بيانات التدريب. بشكل أساسي تتعرض النماذج الواسعة السطحية بسهولة لحدوث فرط التلبيق، بينما النماذج الضيقة العميقة فهي معرضة لحدوث خفض التلبيق[4] underfit. عندما تكون إحدى الطبقات صغيرة وتسمى غالبًا bottleneck عنق الزجاجة، يتدهور الأداء بشكل كبير خاصة عندما تكون طبقة عنق الزجاجة قريبة من طبقة الدخل. وإذا كانت كل طبقة تحتوي على نفس عدد العصبونات، فإن إضافة المزيد من الطبقات قد تؤدي أيضًا إلى تحويل النموذج من حالة فرط التلبيق إلى خفض التلبيق، وذلك لأن الطبقات الإضافية تفرض قيودًا إضافية على موسطات النموذج. للحصول على شبكة جيدة نقوم بتحسين عدد العصبونات في كل طبقة على شبكة ذات طبقة واحدة خفية أولاً، ثم نقوم بإضافة المزيد من الطبقات بنفس الحجم. في مهام التعرف على الكلام، يستخدم عادة 5–7 طبقات DNNs مع 1000–3000 عصبون في كل طبقة.

4.2.2. أصناف شبكات التعلم العميق

يمكن تصنيف شبكات التعلم العميق استناداً لاستخدام بنى وتقنيات مختلفة كالتركيب أو التوليد، والتعرف أو التصنيف إلى الأصناف التالية [22]:

❖ **شبكات التعلم العميق للتعلم غير الموجه unsupervised or generative learning**

تهدف هذه الشبكات للحصول على ارتباط عال في البيانات بغرض تحليل أو تركيب الأنماط عند عدم توفر معلومات صنف الفئة الهدف target class، ويمثل هذا النوع من شبكات التعلم العميق في الأدبيات التعلم غير الموجه للتمثيل أو السمات، مثل آلة بولتزمان المقيدة Restricted Boltzmann Machine (RBMs)، شبكات المعتقدات العميقة Deep Belief Networks (DBNs)، المرمّزات الآلية Regularized Autoencoders.

---

[3] فرط التلبيق Overfitting هي الحالة التي يكون فيها أداء النموذج جيداً مع بيانات التدريب، بينما يكون سيئاً مع بيانات الاختبار فهو غير قادر على التعميم مع بيانات لم يتم تدريبه عليها سابقاً.

[4] خفض التلبيق Underfitting هي الحالة التي يكون فيها أداء النموذج ضعيفاً مع بيانات التدريب بسبب عدم قدرته على إيجاد العلاقة التي تربط بين دخله وخرجه بشكل مناسب.

❖ **شبكات التعلم العميق للتعلم الموجه supervised learning**

في هذا النوع يكون الهدف إعطاء تمييز مباشر للأنماط في مسائل التصنيف، كإعطاء توزيعات احتمالية لاحقة للأصناف، ويكون وسم البيانات (label) معروفاً. تسمى هذه الشبكات شبكات عميقة تمييزية مثل شبكات التعلم العميق Deep Neural Networks (DNNs)، والشبكات العصبونية التكرارية Recurrent Neural Networks (RNNs)، والشبكات العصبونية التلافيفية Convolutional Neural Networks (CNNs).

❖ **شبكات التعلم العميقة الهجينة hybrid**

تهدف لتمييز الأنماط بالاعتماد على كل من الصنفين السابقين بحيث يكون القرار في أحدهما مدعوماً بخرج النوع الآخر، وغالباً ما يتم استخدام التعلم غير الموجه (النماذج التوليدية) في تحسين عملية التمييز للتعلم الموجه.

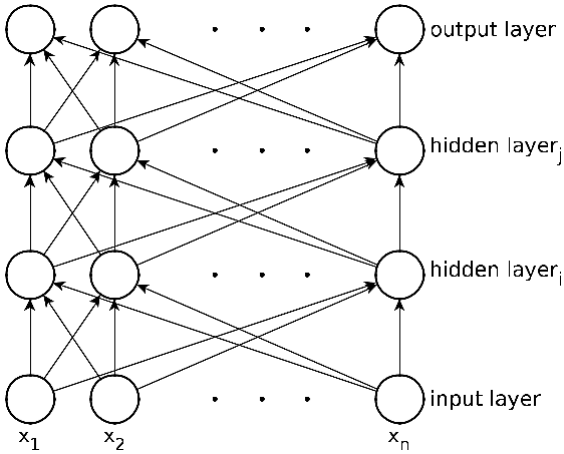
تُعدّ شبكات التعلم العميق الموجه أكثر فاعلية عند التدريب والاختبار، وأكثر مرونة في البناء وأكثر ملاءمة للنظم المعقدة، من ناحية أخرى تعد شبكات التعلم غير الموجه وخاصة الاحتمالية، أسهل في التفسير التركيب، وأسهل لتضمين معرفة المجال والتعامل مع الشك، لكنها تصبح مستعصية عند الاستدلال والتعلم للأنظمة المعقدة.

سنتحدث بشكل موجز عن شبكات التعلم العميق التي تهم في بحثنا.

5.2.2. شبكات التعلم العميق Deep Neural Networks (DNNs)

يسمى هذا النوع من الشبكات multilayer perceptrons (MLPs)، أو الشبكات العصبونية ذات التغذية الأمامية feedforward neural networks، وتمثل نماذج التعلم العميق الجوهرية. تتكون هذه الشبكة من عدة طبقات خفية تتدفق المعلومات فيها من طبقة الدخل لطبقة الخرج في الاتجاه الأمامي فقط forward، ولا توجد وصلات راجعة يتم فيها إعادة تغذية النموذج بمخرجاته. تعد هذه الشبكات نقطة انطلاق مفاهيمية لباقي أنواع الشبكات وتستخدم في مهام التصنيف والتنبؤ.

يبين (الشكل 5) شبكة عصبونية متعددة الطبقات بتغذية أمامية لها طبقة دخل وطبقتين خفيتين وطبقة خرج واحدة [23].



الشكل 5– شبكة تعلم عميق بطبقة دخل وطبقة خرج وطبقتين خفيتين

الهدف في شبكات التغذية الأمامية إيجاد تقريب لتابع ما f، فمثلاً بفرض لدينا مسألة تصنيف فيها دخل x يقابله خرج y، تعرف شبكات التغذية الأمامية تابع تقابل $y = f(x;\theta)$ يتعلم قيمة الموسطات θ التي تعطي أفضل تقريب لهذا التابع[16]. تتشكل شبكة التغذية الأمامية بارتباط مجموعة توابع، ويحدد النموذج المرتبط بالشبكة كيفية ربط هذه التوابع بعضها البعض، على سبيل المثال بفرض لدينا ثلاثة توابع في بنية تسلسلية، عندها يصبح النموذج $f(x) = f^3(f^2(f^1(x)))$; حيث f^1 تابع الطبقة الأولى، f^2 تابع الطبقة الثانية، وهكذا..

لتوضيح طريقة عمل الشبكة رياضياً [17]، لدينا شبكة عصبونية بتغذية أمامية، يمثل فيها الدخل بالمصفوفة $X \in R_{n\times d}$ تحوي n مثالاً للتدريب، لكل مثال d مدخل أو سمة. وبفرض لدينا طبقة خفية تحوي h عقدة. خرج الطبقة $H \in R_{n\times h}$ حيث H شعاع الطبقة الخفية. بفرض الطبقة الخفية ترتبط بطبقة الخرج بشكل كامل وبفرض أوزان الطبقة الخفية $W^{(1)} \in R_{d\times h}$، والانحياز $b^{(1)} \in R_{1\times h}$، وأوزان طبقة الخرج $W^{(2)} \in R_{h\times q}$ ، والانحياز $b^{(2)} \in R_{1\times q}$، عندها يتم حساب خرج الشبكة $O \in R_{n\times q}$ كما يلي:

$$H = \sigma\left(XW^{(1)} + b^{(1)}\right)$$

$$O = \sigma\left(HW^{(2)} + b^{(2)}\right)$$

يمثل σ تابعاً غير خطياً يسمى تابع التفعيل.

وفي حال وجود أكثر من طبقة خفية تصبح المعادلة:

$$H^{(n)} = \sigma_n(H^{(n-1)}w^{(n)} + b^{(n)})$$

حيث $H^{(n-1)}$ يمثل خرج الطبقة السابقة.

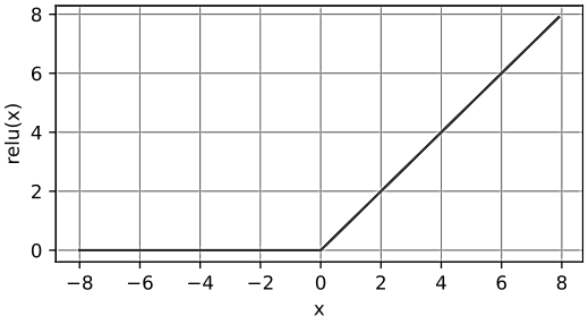
6.2.2. توابع التفعيل Activation Functions

هي توابع غير خطية، تهدف لإضافة خاصية اللاخطية للشبكة العصبونية مما يسمح بتعلم أنماط معقدة في البيانات. يأخذ التابع إشارة خرج الخلية السابقة ويحولها إلى شكل ما قبل انتقالها للخلية التالية، ليحدد ما إذا كان يجب تنشيط العصبون أم لا. فيما يلي أهم توابع التفعيل:

❖ **تابع التفعيل ReLU Function (Rectified Linear Unit)**

التابع الأكثر شيوعًا، نظرًا لبساطته وأدائه الجيد في مجموعة متنوعة من المهام. يوفر ReLU عملية تحويل غير خطية بسيطة للغاية. بفرض لدينا x ، يعرف تابع ReLU على أنه القيمة العظمى لكل من x و 0 كما في (الشكل 6)

$$\mathrm{ReLU}(x) = \max(x,0)$$



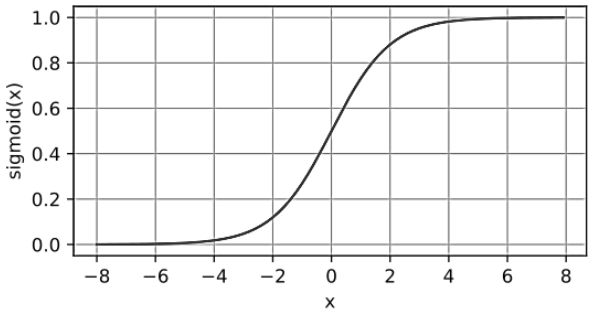
الشكل 6- تابع التفعيل ReLU

يحتفظ هذا التابع بالقيم الموجبة فقط ويتخلص من القيم السالبة بجعل قيمتها صفراً. سبب استخدام هذا التابع هو أن مشتقاته تعمل بشكل جيد فهي إما أن تتلاشى أو تمرر القيمة كما هي، وهذا يجعل عملية التحسين أفضل ويخفف من مشكلة تلاشي التدرجات vanishing gradients التي تعاني منها معظم الشبكات العصبونية.

❖ **تابع التفعيل Sigmoid**

يعمل هذا التابع على تحويل قيم الدخل من مجال قيم الأعداد الحقيقية R للمجال بين 0 و 1، لهذا السبب يطلق على هذا التابع اسم الساحق squashing لأنه يسحق قيم الدخل في المجال $(-inf, inf)$ لتصبح في المجال (0,1) كما في (الشكل 7)

$$\mathrm{sigmoid}(x) = \frac{1}{1 + \exp(-x)}$$



الشكل 7– تابع التفعيل Sigmoid

يأخذ مشتق هذا التابع القيم بين 0.25 (عندما يكون الدخل صفراً) و0 (عندما تبتعد قيم الدخل عن 0).

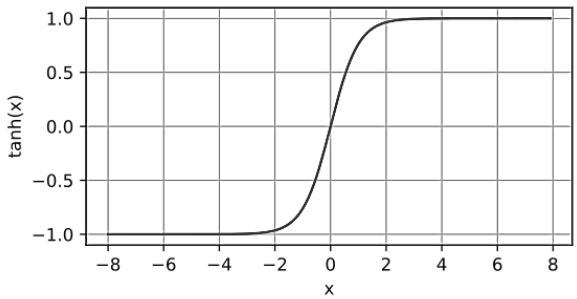
يستخدم هذا التابع على نطاق واسع مع وحدات الخرج، عندما يمثل الخرج احتمال لمسائل التصنيف الثنائي مثلاً، بينما يتم استخدام ReLU مع معظم الطبقات الخفية لسهولته وبساطته في التدريب.

❖ تابع التفعيل Tanh

يسمى أيضاً تابع المماس الزائدي hyperbolic tangent.

يعمل هذا التابع بشكل مشابه لتابع sigmoid فيخفض الدخل ويحول قيمه للمجال بين −1 و1 كما في (الشكل 8)، بينما يأخذ مشتقه القيم بين 0 (عندما تبتعد قيم الدخل عن 0) و1 (عندما يكون الدخل صفرياً).

$$\tanh(z) = \frac{\mathrm{e}^{z} - e^{-z}}{\mathrm{e}^{z} + e^{-z}}$$



الشكل 8– تابع التفعيل Tanh

7.2.2. الذاكرة طويلة قصيرة الأمد Long Short Term Memory (LSTM)

يقتصر استخدام الشبكات العصبونية ذات التغذية الأمامية على مهام التصنيف الثابت [23]، فهي تعتمد تقابلاً ثابتاً بين الدخل والخرج. لذا نحتاج إلى ما يسمى بالمصنف الديناميكي لنمذجة المهام المرتبطة بالزمن.

يمكننا تطوير الشبكات العصبونيّة ذات التغذية الأمامية لتصبح ديناميكية. وذلك بإعادة إدخال الإشارات من الخطوات الزمنية السابقة إلى الشبكة. وتسمى هذه الشبكات ذات الوصلات المتكررة بالشبكات العصبونيّة المتكررة (RNN). تمتلك شبكات RNN حالة داخلية تحتفظ فيها بالمعلومات في كل خطوة زمنية. تقتصر قدرة ذاكرة RNNs على حفظ ما يقارب عشر خطوات زمنية، وذلك بسبب أن إشارة التغذية الخلفية إما أن تتلاشى أو تنفجر. تمت معالجة هذه المشكلة من خلال الشبكات العصبونيّة المتكررة للذاكرة طويلة قصيرة الأمد (LSTM-RNN). إن شبكات LSTM قادرة على تعلم أكثر من 1000 خطوة زمنية، وذلك حسب تعقيد الشبكة المبنية.

❖ **بنية شبكة LSTM**

تم تحسين قدرة التذكر للخلية العصبونية المتكررة عن طريق إدخال مفهوم "بوابة" في الخلية، لتصبح الخلية مكونة من بوبة إدخال input gate، وبوابة إخراج output gate، وبوبة نسيان forget gate. ويوضح (الشكل 9) بنية خلية ذاكرة LSTM [24].

تحدد بوابة الإدخال المعلومات الجديدة التي سيتم تخزينها في الخلية، وتحدد بوابة الإخراج المعلومات التي سيتم إخراجها بالاعتماد على حالة الخلية، بينما تحدد بوابة النسيان المعلومات التي سيتم التخلص منها عن حالة الخلية. عندما تكون قيمة بوابة النسيان 1 فإنها تحتفظ بالمعلومات على عكس ذلك يتم التخلص من جميع المعلومات عند القيمة 0.

تتم في كل خلية العمليات التالية

$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\
f_t &= \sigma\left(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}\right) \\
g_t &= \tanh\left(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}\right) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

حيث:

x_t دخل الخلية في اللحظة t

h_t الحالة الخفية في اللحظة t

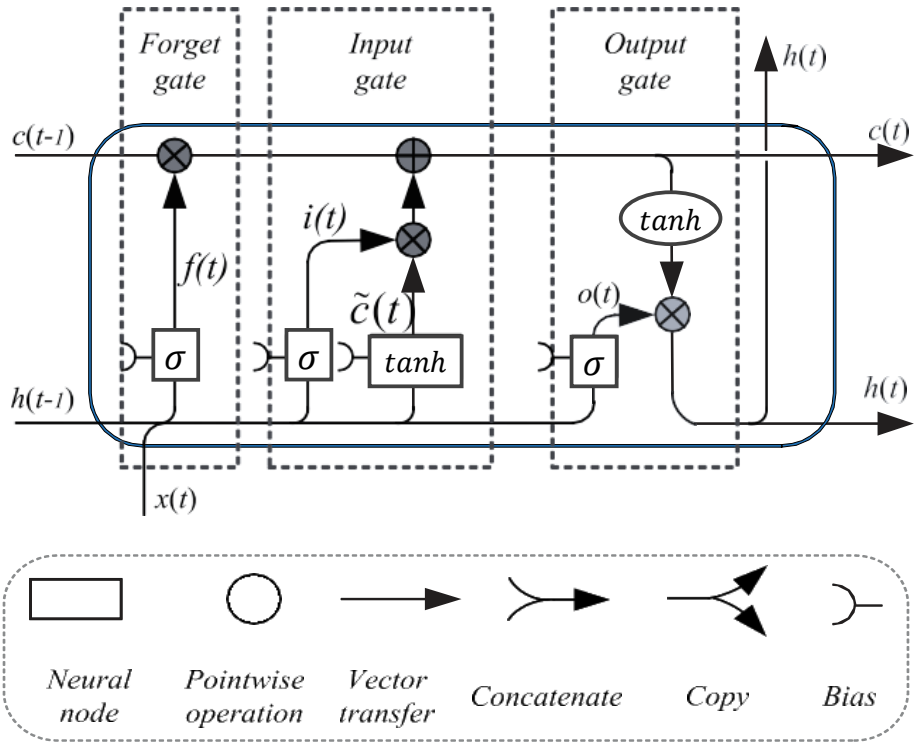
c_t حالة الخلية في اللحظة t

i_t , f_t , g_t , o_t قيمة بوابة الدخل والنسيان والخلية والخرج على الترتيب في اللحظة t

W_{ij} الوزن الذي يربط بين الوحدتين i,j ، b الانحياز

σ تابع sigmoid

$\odot$ جداء هارمارد Hadamard[5]



الشكل 9– بنية خلية ذاكرة LSTM

سنوضح سبب استخدام هذا النوع من الشبكات في بحثنا في فصل المقاربة المقترحة.

[5] يسمى أيضاً element-wise https://en.wikipedia.org/wiki/Hadamard_product_(matrices)

3.2. خاتمة

تحدثنا في هذا الفصل عن علم الأصوات وتناولنا بشكل خاص تصنيف الأصوات وفق مخارج الحروف وصفاتها بالنسبة للّغة العربية، ثم تكلمنا عن معالجة الإشارة الصوتية مع أبرز السمات التي يتم استخلاصها من إشارة الكلام. تناولنا بعد ذلك شبكات التعلم العميق بشكل عام وأنواعها وتحدثنا بشكل خاص عن شبكات التعلم العميق DNNs والشبكات ذات الذاكرة طويلة قصيرة الأمد LSTM. سننتقل في الفصل التالي للحديث عن الدراسات المرجعية للأعمال ذات الصلة بمشكلة البحث التي نحاول حلها.

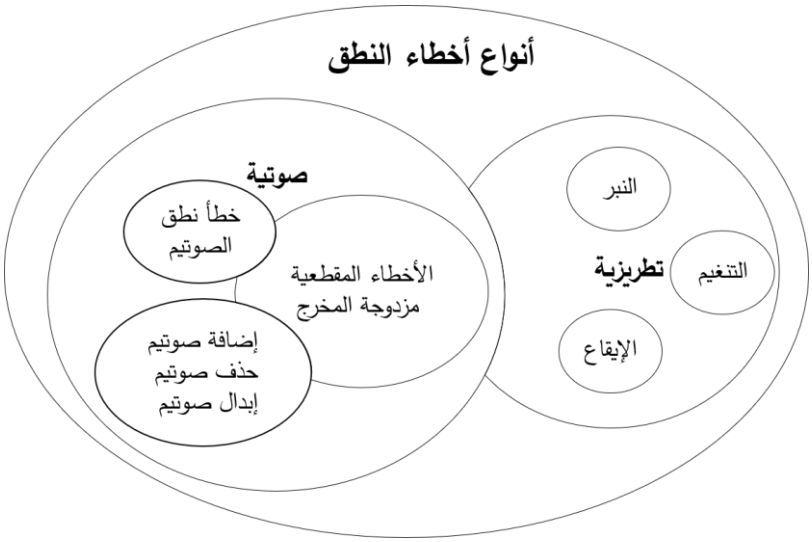
الفصل الثالث: الدراسة المرجعية

1.3. مقدمة

يشير مصطلح النّطق (Pronunciation) للطريقة التي يتم فيها اللفظ وفق كيفية معينة بلغة ما، ويعد اللبنة الأساسية في بناء المعنى اللغوي [25] لأن اختلاف لفظ الأصوات يؤدي لإحداث تمايز دلالي في المعنى، فكلمتي "صار" و "سار" لكل منهما معنى معين مختلف عن الآخر لاختلاف لفظ الصوتيم (ص) عن (س) فيهما.

يتضمن النّطق تفاعل مجموعة من العوامل الإدراكية والتعبيرية، ويتم وصفها من خلال مجموعة من السمات (القِطَعيّة segmental، الصوتية voice، التطريزية prosodic، ...) التي تُمكّن من إعطاء مؤشر على جودة النطق، إلا أنه يصعب قياس خطأ النّطق بشكل كمي، إذ لا يوجد ما يعرّف بدقة صحة النّطق أو عدمه، لذا يتم قياس مدى وضوح النّطق وقربه من نطق متحدث أصلي للغة، ويمكن تصنيف أخطاء النّطق بشكل عام لصوتية phonemic وتطريزية prosodic [26]. تعد الأخطاء الصوتية الأهم والأكثر حساسية فهي تشمل حالات استبدال صوتيم بآخر أو حذفه أو إضافة صوتيم على الكلمة المنطوقة، وعلى نطاق أصغر نجد الأخطاء المتعلقة بنطق الصوتيم نفسه بطريقة مختلفة عن النّطق الصحيح وذلك بتغيير أحد سماته الصوتية. أما الأخطاء التطريزية فترتبط بالسمات الصوتية المصاحبة للكلام مثل النبر stress والتنغيم intonation والإيقاع rhythm والمدة duration.

إن أخطاء النّطق مرتبطة ببعضها البعض كما يبين (الشكل 10)، وهذا يجعل النّطق مسألة متعددة الأبعاد من الصعب تحديدها من خلال منهجية واحدة، لذا فإن بناء نظام ناجح مساعد على تصحيح النّطق يتطلب استخدام عدة تقنيات مختلفة.



الشكل 10– أنواع أخطاء النّطق

ركزت معظم الأبحاث على أخطاء النّطق الصوتي phonemic [1] كونها تحدد بشكل أدق خطأ النّطق في الكلام وتعطي المتعلم دلالة واضحة للخطأ المرتكب، حيث ترتبط هذه الأخطاء باللغة المدروسة وتختلف من لغة لأخرى كاختلاف عدد الصوتيمات وطبيعتها [27]، بينما نجد أن الأخطاء التطريزية prosodic تشترك فيها معظم اللغات إلا أن أهمية السمة الصوتية بين اللغات هي التي تختلف، وبالتالي فهي لا تفيد متعلم اللغة بشكل كبير على تصحيح نطقه.

تم العمل وفق عدة منهجيات لكشف الأخطاء الصوتية في النّطق ويمكن تصنيف هذه المنهجيات حسب عدة عوامل [1] كاعتمادها على وجود نظام تعرف آلي على الكلام، والحاجة لوجود معرفة باللغة الأصلية ولغة المتعلم إضافة لقدرة المنهجية على كشف وجود خطأ في النّطق وتحديد طبيعته ومكانه.

نستعرض في هذا البحث المنهجيات المتبعة في كشف أخطاء النّطق وفق العوامل السابقة، ثم نتناول أهم الأبحاث فيها بشكل مفصل أكثر.

2.3. المنهجيات العامة المتبعة

1.2.3. المنهجيات المعتمدة على درجة الثقة (Confidence score based)

يتم في هذه المنهجيات حساب درجة للثقة تعبر عن مدى قرب اللفظ المنطوق من اللفظ الصحيح، ثم تتم مقارنة النتيجة مع عتبة محددة لأخذ القرار فيما إذا كان النّطق صحيحاً أم لا. بدأ العمل بهذه المقاربة سنة 1997 وتم اعتماد عدة مقاييس للمقارنة منها HMM-based , HMM-based log-likelihood log posterior [28]. ويعد مقياس صحة النّطق Goodness Of Pronunciation (GOP)[6] الأكثر استخداماً [29] حيث يعتمد هذا المقياس على حساب نسبة احتمال likelihood ratio توافق الصوت المنطوق مع الصوت الصحيح. وقد تم العمل على تحسين جودة النموذج الصوتي المستخدم باعتماد شبكات التعلم العميق DBN-HMM بدلاً من النموذج الغوصي GMM-HMM [30] [6]. إن مقياس GOP حساس جداً لجودة النموذج الصوتي [29]، إضافة إلى أن هذه المنهجيات تستخدم وسم البيانات ذات النّطق الخاطئ في مجموعة بيانات التدريب لهدف محدد كتحديد عتبة القرار فقط، مما يستدعي وجود منهجيات أخرى تستخلص قدراً أكبر من المعلومات المفيدة من البيانات الموسومة [31]، كما أن عتبة القرار يتم تحديدها بالاعتماد على بيانات التدريب فقط، مما جعل من الصعب اعتماد هذه المنهجية للتعميم من أجل طيف واسع من أخطاء النّطق. وعلى الرغم من أن هذه المنهجية يمكنها الكشف عن الأخطاء لكنها غير قادرة على تشخيص نوع الخطأ الذي ارتكبه المتعلم.

---

[6] مقياس صحة النّطق GOP مقياس يستخدم لتقييم جودة اللفظ للمتحدثين غير الأصليين للغة.

2.2.3. المنهجيات المعتمدة على القواعد (Rule based)

تتطلب هذه المنهجيات معرفة مسبقة بقواعد النّطق الخاطئ وتتميز بقدرتها على تحديد مكان الخطأ الذي قام به المتكلم ونوعه، إلا أنها محصورة بالأخطاء المعرفة مسبقاً ضمن القواعد، إذ لا يمكن للنظام التّعرف على أخطاء جديدة غير معرّفة مسبقاً وتحديد مكان الخطأ. تُبنى قواعد النّطق الخاطئ يدوياً من قبل خبراء لغويين [32] [33]. كما يُبنى نموذج صوتي للنطق الصحيح، ثم تُطبّق القواعد لتوليد نماذج للنطق الخاطئ، بعد ذلك تتم المقارنة بين خرج النموذجين عند اختبار النّطق للمتعلّم لتحديد صحة النّطق. تم اقتراح تشكيل عناقيد لقواعد النّطق باستخدام شجرة قرار، وإسناد عتبات مختلفة للخطأ لكل عنقود، مما يجعل القرار بشأن النّطق أقل صرامة ومماثل أكثر للقرار البشري [34]. يتطلب بناء قواعد النّطق وجود خبرة لغوية إضافة إلى أن هذه القواعد قد لا تغطي كل الحالات، لذا تم تحسين هذه الطرق باستنتاج قواعد النّطق الخاطئ آلياً باستخدام مجموعات معطيات لمتعلم اللغة الثانية L2[7] [35]. تعتمد معظم هذه الطرق على بناء نموذج صوتي باستخدام نموذج ماركوف HMM, وقد جرى العمل لاحقاً على تطويرها باستخدام نماذج صوتية محسنة تعتمد على GMM-HMM وبعدها DNN-HMM [36]، كما تم العمل على المكاملة بين هذه المنهجية والمنهجية السابقة المعتمدة على درجة الثقة للحصول على نتائج أفضل[37].

3.2.3. المنهجيات المعتمدة على المصنفات (Classifier based)

تعتمد هذه المنهجيات على تصنيف الوحدات الصوتية باستخدام مصنفات مختلفة كمصنف آلة شعاع الدعم Support Vector Machine (SVM) أو أشجار التصنيف Decision Trees أو الشبكات العصبونية Neural Networks. تُستخلص سمات مختلفة من الإشارات الصوتية في المرحلة الأولى، لتكون دخلاً للمصنفات في المرحلة التالية. في هذا العمل [38] تمت المقارنة بين منهجيتي التصنيف باستخدام linear-discriminant analysis (LDA) ومنهجية درجة الثقة Goodness Of Pronunciation لتصنيف زوج من الأصوات في اللغة الهولندية وأعطت منهجية التصنيف LDA أفضل النتائج. وقد أثبتت مصنفات SVM فعاليتها في العديد من الأبحاث[39] [31]. وفي [40] اقتُرِح استخدام مصنف a Neural Network (NN) based Logistic Regression بالاعتماد على الشبكات العصبونية وجرت مقارنته مع مصنف SVM ومنهجية GOP حيث أعطى نتائج أفضل في الحالتين، كما استُخدمت شبكات التعلم العميق لتحسين جودة النموذج الصوتي المستخدم في استخلاص سمات الوحدات الصوتية قبل تدريب المصنف. وبصورة عامة فإن هذه المنهجية تعتبر أفضل من كل من المنهجيتين السابقتين، إلا أنها بالمقابل تحتاج كمية كافية من بيانات التدريب.

---

[7] L2 يشير هذا الاختصار لمتعلم اللغة الثانية second-language أي اللغة غير الأصلية للمتحدث، التي قام بتعلمها لاحقاً.

4.2.3. المنهجيات المعتمدة على التعلم العميق (Deep Neural network based)

في هذه المنهجيات يُستخدم نوع خاص من الشبكات العصبونية وهي شبكات التعلم العميق، لنمذجة التجريدات العالية المستوى في البيانات، واستخلاص سمات قادرة على تحسين دقّة كشف أخطاء النّطق كما في الأبحاث [41] ,[2]. في بعض الأبحاث تم استخدام الشبكات العصبونية التلفيفية Deep Convolutional Neural Network (CNN) لاستخلاص السمات الممثّلة للإشارات الصوتية من الصور[4]. وبهدف تحسين كشف خطأ النّطق بالاستفادة من شبكات تعلّم مدرّبة مسبقاً، تم العمل وفق طريقة التعلّم المنقول transfer learning[8]، فقد أثبتت التّجارب أن هذه النماذج المدرّبة يمكن تعميمها على مهام عديدة، كاستخدام نموذج لغة مدرب كمرحلة أولية في التصنيف في لغة أخرى [42]، وهذا يوفر الوقت ويساعد بشكل كبير في حال وجود موارد لغوية محدودة [43]. كما تم استخدام شبكات التعلم العميق في بعض الأبحاث لتحسين دقّة النموذج الصوتي كما ذكرنا سابقاً.

بالنسبة للسّمات المستخدمة في الأبحاث فتنقسم إلى سمات صوتية يتم استخلاصها من البيانات مثل (MFCC, Mel spectrogram , statistical features...) وتعتبر معاملات ميل التردّدية MFCC مع مشتقّاتها الزمنية (معاملات دلتا ودلتا دلتا) من الدرجة الأولى والثانية، الأكثر استخداماً وجدوى كونها تحاكي استجابة النظام السمعي للإنسان، حيث أثبتت فعاليّتها في العديد من الأبحاث وتم استخدامها في عدة منهجيات مختلفة، أما السمات النّطقيّة فتم توظيفها حديثاً في مجال المعالجة الآلية للصوت وتسمى واصفات النّطق الكلامية، وهي خصائص صوتية تصف آليّة نطق الحروف ومخارجها، وتستطيع أن تميّز صوتاً عن صوت آخر [9]، وهذا يوفّر لمتعلّم اللغة معلومات أدق عن صحة نطقه ويساعده في تحديد موضع الخطأ وتصحيحه.

نستعرض في الفقرة التالية بمزيد من التفصيل أهم الأعمال التي تمت وفق المنهجيات السابقة

3.3. الأعمال السابقة

- قام الباحثان M. Shahin and B. Ahmed في [44] باقتراح منهجيّة للتحقق من صحة نطق اللّغة الإنجليزية بوصفها طريقة للكشف عن وجود شذوذ في الكلام، حيث تم تدريب نموذج صوتي خاص بالوحدات الصوتية (الصوتيمات) قادر على تحديد وجود خلل في النّطق، من خلال بيانات تدريب خاصة بالنّطق الصحيح فقط دون الحاجة لبيانات تدريب خاصة بالنّطق الخاطئ، تم استخدام مصنف أحادي One-Class SVM لكل صوتيم يعتمد على واصفات الكلام المتعلقة بمخارج وصفات الحروف

[8] التعلم المنقول هو أحد تقنيات تعلم الآلة التي تهدف لتحسين عملية التعلم في مهمة جديدة من خلال نقل المعرفة من مهمة ذات صلة تم تعلمها سابقاً.

(manners and places of articulation) التي تم تحديدها باستخدام شبكات التعلّم العميق (DNN). وقد بلغت دقة تمييز الواصفات 91.5% ± 2.5%، بينما بلغت نسبة كشف خطأ الصوامت consonants باستخدام مقياس F1[9] (F-measure) ما يقارب 0.88 ± 0.05 وللصوائت vowels 0.86 ± 0.04. تم تدريب النظام على قاعدتي البيانات WSJ0 [10] و TIMIT[11] وتحوي كل منهما على بيانات صوتية لمتحدثين أصليين للغة الإنجليزية. تم تقييم النظام بالاعتماد على بيانات صوتية تحتوي على أخطاء في النّطق وكانت دقة مقياس F1 هي 0.84 من أجل مجموعة البيانات الأصلية الحاوية على أخطاء مصطنعة. بينما سجل المقياس دقة 0.82 من أجل مجموعة البيانات المنطوقة بلكنة غير أصلية (The GMU foreign-accented speech)، و0.80 من أجل مجموعة بيانات صوتية للأطفال تحوي على اضطرابات في النّطق [12].

- تم العمل في هذا البحث [45] على تحسين أداء عملية تحديد النّطق الخاطئ من قبل متعلمي اللغة باستخدام واصفات الكلام المتعلقة بمخارج الحروف وصفاتها، مع شبكات التعلم العميق والذواكر طويلة قصيرة الأمد LSTM. تم تدريب النظام باستخدام مجموعة تدريب للمتحدثين الأصليين للغة الصينية [13] والتي تحوي تسجيلات صوتية بطول 100 ساعة تقريباً ل 150 متحدث إضافة لجزء من مجموعة بيانات لمتحدثين غير أصليين iCALL[14] وذلك لبناء النموذج الصوتي، كما تم اختبار النظام من خلال جزء من بيانات المجموعة iCALL تحوي على تسجيلات صوتية ل 30 متحدث من جنسيات مختلفة. يتم في المرحلة الأولى بعد تدريب النموذج الصوتي إجراء تقابل بين الصوتيمات وواصفات الكلام، وتم استخدام خمس تصنيفات أساسية وتدريب شبكة تعلم عميق منفصلة لكل منها لتمييز وجود الواصفة في الإشارة الصوتية، في المرحلة الثانية تم دمج واصفات الكلام من المرحلة السابقة مع سمات الصوتيمات لتدريب شبكة LSTM تعمل على تمييز النّطق الخاطئ، حيث تم الحصول على نسبة false acceptance rate (FAR) مساوية ل 8.65 ونسبة false rejection rate (FRR) مساوية ل 3.09 وهي أفضل من النتائج التي تم الحصول عليها في دراسة سابقة مشابهة تم استخدام الشبكات العصبونية ANN فيها.

---

[9] F-measure مقياس لتقييم النموذج يجمع بين الدقة precision والاسترجاع recall ويستخدم عند تقييم نظم التصنيف الثنائي.

[10] مجموعة معطيات Wall Street Journal (WSJ0) القياسية تتكون من 101 متحدثًا و13048 جملة (~21 ساعة لكامل المجموعة).

[11] مجموعة معطياتTIMIT (Texas Instruments Massachusetts Institute of Technology).

[12] مجموعة معطيات للكلام المضطرب تم جمعها من الأطفال المصابين بتعذر الأداء النطقي في مرحلة الطفولة Childhood Apraxia of Speech (CAS)

[13] مجموعة الكلام للمتحدثين الأصليين للغة الصينية للمشروع الوطني الصيني للتكنولوجيا الفائقة 863 لتطوير نظام LVCSR .

[14] iCALL مجموعة كلام للمتحدثين غير الأصليين للغة الصينية.

- تم في هذا العمل [6] تحسين أداء عملية تحديد النّطق الخاطئ للغة الصينية وفق المنهجية المعتمدة على درجة الثقة ووفق منهجية المصنفات أيضاً. تم بناء نموذج صوتي باستخدام مقاربة GMM-HMM ثم تم تحسين هذا النموذج باستخدام شبكات التعلم العميقDNN ، وبعد الحصول على النموذج الصوتي تم حساب معيارGOP لتقييم صحة النّطق والحصول على دقة تحديد النّطق الخاطئ بما يقارب 62.8% للغة الصينية من أجل 20 صوتيم يتم نطقهم بشكل خاطئ غالباً، و44% للغة الإنجليزية من أجل 15 صوتيم تحدث فيهم أخطاء النطق أكثر من غيرهم أيضاً. أما بالنسبة لمنهجية التصنيف تم استخدام مصنف Logistic Regression (LR) معتمد على التعلّم المنقول transfer learning لتمييز وجود خطأ النّطق بعد أن يتم تصنيف الصوتيمات في مرحلة سابقة، وقد بلغت الدقة في هذه المنهجية 75.5 % للغة الصينية و 69.2% للغة الإنجليزية من أجل نفس مجموعة الصوتيمات السابقة، وأعطت نتيجة أفضل من المنهجية المعتمدة على درجة الثقة. بالنسبة لبيانات التدريب تم استخدام مجموعة تدريب لناطقين أصليين للّغة ثم اختبار نظام كشف النّطق الخاطئ باستخدام بيانات تدريب لمجموعة من المتعلّمين، لكل من اللغتين الإنجليزية والصينية.

- عمل الباحثون في [46] على تحديد أخطاء النّطق في اللغة العربية وقاموا بتسجيل بيانات صوتية للأحرف العربية الساكنة (28 حرف)، وذلك من قبل 200 متحدث بأعمار مختلفة وقدرات نطق متفاوتة، وقد قام مجموعة من الخبراء في النّطق بتحديد صحة النّطق في هذه البيانات. كما تم تقسيم الحروف إلى مجموعتين وفق تشابه نطق النهاية الصوتية للحروف، ثم تدريب كل مجموعة على مصنف خاص لاختبار صحة النّطق مع تقسيم المقاطع الصوتية لعشرة أجزاء متساوية segments، واستخدام بعض هذه الأجزاء لاستخلاص السمات في المجموعة الأولى لتكون قادرة على تمييز الأصوات بشكل أكبر، واستخدام جميع الأجزاء من أجل سمات المجموعة الثانية، ثم استخدام الشبكات العصبونية في عملية التدريب وبلغت دقة التصنيف 82.27% لهذا النظام.

- في العمل [4] المختص باللغة العربية أيضاً تم اقتراح استخدام الشبكات العصبونية التلفيفية deep convolutional neural network (CNN) بالطريقة التقليدية وطريقة التعلّم المنقول لاستخلاص السمات، ثم استخدام أحد المصنفات K-Nearest Neighbor (KNN)، support vector (SVM) machines، Neural Networks (NN) لتحديد النّطق الخاطئ. وبلغت دقة التصنيف عند استخدام سمات CNN 91.7%، بينما تم الحصول على دقة 92.2% من أجل طريقة التعلّم المنقول. تم الاعتماد على بيانات مسجلة من قبل 400 متحدث يتعلمون اللغة العربية كلغة ثانية وقام مجموعة من الخبراء بوسم هذه البيانات كصحيحة أو خاطئة. جرت عملية التدريب على 6 مجموعات للأصوات المتشابهة ومجموعة منفصلة لباقي الحروف.

- جرى العمل في البحث [11] بشكل خاص على اختبار واصفات الكلام المتعلقة بمخارج وصفات الحروف من أجل اللغة العربية. وقد تم تدريب شبكة تعلّم عميق ثنائية من أجل كل واصفة. جرى استخدام بيانات تدريب صوتية للقرآن الكريم بطول 90 ساعة، وتراوحت دقة التصنيف في هذا النظام بين 76% و95% من أجل الواصفات المختلفة. تمت عملية معالجة البيانات من أجل تقطيعها وتقسيمها من مستوى الآيات لمستوى الوحدات الصوتية باستخدام المحاذاة القسرية force alignment، ثم إسناد كل وحدة صوتية للميزة التي تتبع لها والقيام بعملية التدريب.

- قام الباحثون في هذا العمل [47] باقتراح نماذج الانتباه attention-based لتعرّف كل من الصوتيمات وواصفات النطق. توفر هذه النماذج ربط بين مرمّز (البيانات الصوتية) ومفكك ترميز (البيانات النصية) لتشكيل محاذاة بين خرج مفكك الترميز ودخل المرمّز. جرى اعتماد نموذج الانتباه LAS[15] وتطبيقه وفق عدة سيناريوهات واختباره على مجموعة معطيات TIMIT. حصل النموذج على أفضل النتائج باعتبار الصوتيمات خرجاً للنموذج المقترح حيث حقق معدل خطأ صوتيم 20.2% لتعرف الصوتيمات. بالنسبة لتعرف وصفات النطق تم اقتراح منهجية التعلم المتعدد المهام لتعلم كل من مهمتي تصنيف الصوتيمات وواصفات النطق وتطبيق نموذج LAS المتعدد المهام الذي يقوم بمشاركة موسطات المرمّز بين المهمتين السابقتين ويعتمد خرج مفكك الترميز (الواصفات المقابلة للصوتيمات) خرجاً للنموذج. وقد حصل هذا النموذج على نسبة تعرف وسطية للواصفات 95.5% على مستوى الإطار الصوتي.

- اقترح الباحثون في [48] استخدام تقنيات كشف الأغراض object detection للتعرف على الصوتيمات. وقد تم اختيار الكاشفين YOLO[16] وCenterNet[17]، حيث يجري تحويل الإشارات الصوتية لصور طيفية spectrogram، ثم تدريب نظام لتحديد الصوتيمات وحدودها الزمنية باستخدام الكاشف المقترح. أعطى النظام بالنسبة للكاشف CenterNet معدل خطأ صوتيم 15.89% على مجموعة معطيات TIMIT. بينما حصل على معدل 16.34% مع كاشف YOLO. في ذات السياق جرى اقتراح استخدام تقنيات كشف الأغراض في العمل [49] للتعرف على تتالي واصفات النّطق. استخدمت هذه الواصفات في مرحلة التعرف على الصوتيمات بمقارنة شعاع الواصفات المتوقع وشعاع

---

[15] نموذج الانتباه Listen, Attend and Spell (LAS) شبكة عصبونية لتحويل الإشارة الصوتية إلى تسلسل الحروف المقابلة لها.

[16] خوارزمية YOLO اختصاراً ل "You Only Look Once " وهي خوارزمية تكشف الأغراض في الصور عن طريق تحديد إحداثيات الصندوق الحدودي bounding box coordinate المحيط بالغرض وحساب احتمال الصنف الذي ينتمي له، وبذلك يتم تحديد الغرض ومكانه في الصورة.

[17] CenterNet خوارزمية لكشف الأغراض في الصور عن طريق تحديد ثلاثية نقاط (نقطة مركزية ونقطتين زاويتين للمستطيل المحيط بالغرض).

واصفات مرجعي وفق نسبة تشابه محددة بينهما. وسمي النظام المقترح ب (AFD-Obj). تتميز هذه الطريقة بعدم الحاجة لتدريب شبكة منفصلة لكل واصفة، وهذا يجعل عملية التدريب أقل تعقيداً وأكثر سرعة. إضافة الى إمكانية الاستفادة من هذه التقنية في تحديد الحدود الزمنية للصوتيمات وترتيبها. تم اختبار النظام على كل من اللغتين العربية (باستخدام مجموعة بيانات KACST Arabic Phonetic Database (KAPD)[50]) والإنجليزية (TIMIT). أعطى الكاشف YOLOv3-tiny[18] نتائج جيدة مقارنة مع منهجيات أخرى بالنسبة لKAPD، حيث بلغت قيمة F-measure لكشف الواصفات 96.5%، وتم الحصول على معدل خطأ للصوتيمات phoneme error rate (PER) 10.84% وفق نسبة تشابه 100% بين شعاع الواصفات المرجعي والشعاع الناتج عن الشبكة. أما بالنسبة ل TIMIT فقد بلغت نسبة دقة كشف الواصفات 95.13%. كما تم اقتراح استخدام الكاشف YOLO لتحديد الصوتيمات والتعرف عليها بنظام آخر (PD-Obj) بشكل مشابه للعمل السابق وحقق النظام المقترح معدل خطأ للصوتيمات 5.63% على مجموعة KAPD.

يبين (الجدول–4) مقارنة بين الأعمال السابقة.

4.3. خاتمة

قمنا في هذا الفصل باستعراض أهم الأعمال في مسألة كشف أخطاء النّطق وأهم المنهجيات التي تستند لها، وتبين أن معظم الأبحاث الحالية توجهت للاعتماد على التعلم العميق كون نتائجه أفضل من الطرق التقليدية ولأنه يلائم وجود كمية كبيرة من البيانات. إن معظم المقاربات السابقة على الرغم من النتائج الجيدة التي وصلت لها إلا أنها تتطلب وجود قدر كافٍ من البيانات المنمّطة الحاوية على مختلف الحالات الممثلة للنطق الخاطئ، وهذا يتطلب جهداً ووقتاً كبيرين في ظل عدم توفر مثل هذه الموارد لبعض اللغات وبشكل خاص لّلغة العربية التي هي محور دراستنا في البحث، وهذا ما جعلنا نتوجه لطريقة أخرى يتم فيها معالجة المسألة على أنها كشف للشذوذ عن الحالة الطّبيعية (النّطق الصحيح)، تتطلب وجود بيانات تدريب النّطق الصحيح فقط. إضافة الى أن تحديد نوع الخطأ غير ممكن دائماً ويتطلب استخدام نوع خاص من السمات، وقد بيّنت الأبحاث أن واصفات النّطق تساعد في هذه المهمة وتعطي معلومات دقيقة عن كل صوت، وتسهل تمييزه عن باقي الأصوات مع توضيح آلية حدوثه إضافة لكشف وجود خطأ فيه وتوفير إمكانية تصحيحه بدقة أكبر، مما دفعنا لدراسة هذه الواصفات بشكل خاص بهدف معرفة أثر استخدامها في أنظمة كشف النّطق الخاطئ.

في الفصل القادم سنتحدث عن منهجية العمل المقترحة.

---

[18] نسخة مصغرة من الكاشف YOLOv3 والذي يمثل النسخة الثالثة من خوارزمية YOLO.

الجدول4 مقارنة بين الأعمال السابقة

العمل	مجموعة البيانات	المنهجية	اللغة	الحاجة لبيانات معلّمة للنطق الخاطئ	كشف وجود الخطأ	تحديد نوع الخطأ	المعيار المعتمد لقياس جودة النظام	
[44]	TIMIT ،WSJ0	DNN من أجل تدريب واصفات النّطق SVM لكشف الشذوذ في النّطق	الإنجليزية	لا	نعم	نعم	DNN speech attribute accuracy	91.5% ± 2.5%
							OCSVM F1 mispronunciation detection	0.86 ± 0.04 vowels 0.88 ± 0.05 consonant
[45]	- non native (iCALL) - native Chinese corpus (100 hours)	DNN من أجل تدريب واصفات النّطق LSTM لكشف خطأ النّطق	الصينية	نعم	نعم	نعم	mispronunciation false acceptance rate 8.65% false rejection rate 3.09%	
[6]	**English**: (non native corpus) LDC95S27 (Native database) **Chinese**: iCALL (non native) native Chinese corpus (~41 h)	Transfer learning based Logistic Regression (LR) classifier	الصينية والإنجليزية	نعم	نعم	لا	mispronunciation detection Accuracy English: 69.2% Chinese: 75:5%	
[46]	مجموعة بيانات مسجلة لصوتيمات منفردة معلّمة يدوياً (200 متحدث)	ANN	العربية	نعم	نعم	لا	mispronunciation detection Accuracy 82.27%	
[4]	مجموعة بيانات مسجلة لصوتيمات منفردة معلّمة يدوياً (400 متحدث)	Transfer learning with CNN	العربية	نعم	نعم	لا	mispronunciation detection Accuracy 92.2%	
[11]	بيانات تدريب لتسجيلات من القرآن الكريم (90 ساعة)	Binary DNN-based classifier	العربية	لا	لا	لا	ضبط تصنيف مخارج الحروف 83%±4.4% ضبط تصنيف صفات الحروف 84%±4.7%	
[49]	KAPD, TIMIT	Object detection using YOLO For attribute detection	العربية والإنجليزية	لا	نعم	نعم	KAPD attribute f-measure 94.5% TIMIT attribute accuracy 95.13%	

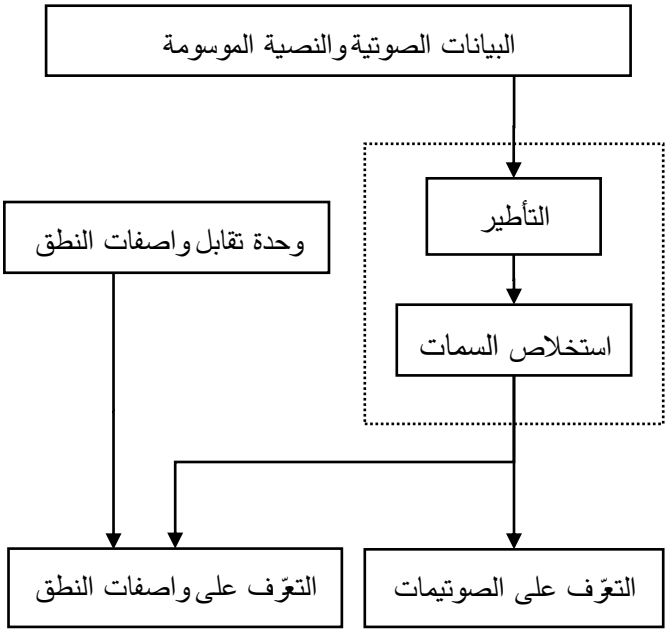
الفصل الرابع: المقاربة المقترحة ومنهجية العمل

1.4. مقدمة

نعرض في هذا الفصل منهجية العمل المتبعة في البحث، نستعرض في البداية المخطط العام للمنهجية. اعتمدنا في النظام المقترح الخطوات التالية: تمر عملية كشف خطأ النطق بدايةً بمرحلة أولى للتعرف على ما جرى لفظه بهدف معرفة فيما إذا كان تم لفظ الصوتيم أم لا، وذلك باعتبار أن متعلم اللغة يقوم بقراءة نص معروف مسبقاً لدى النظام، أي نفترض وجود نص مرجعي يجري تعلمه والتدرب على نطقه. في المرحلة التالية يجري تحليل الصوت بشكل دقيق من خلال تعرّف مجموعة سمات تعطي معلومات كافية عن اللفظ وبالتالي تقدم صورة واضحة عن طبيعة خطأ النطق في حال وجوده، ليتم لاحقاً الاستفادة من هذه المعلومات في توجيه المتعلم لمعرفة الخطأ المنطوق ولإرشاده لكيفية تصحيح الخطأ بالشكل المناسب. وقد وجدنا حسب الدراسات السابقة أهمية استخدام واصفات الكلام المتعلقة بمخارج الحروف وصفاتها للقيام بتحليل الصوت ومعرفة خصائصه وتمييز صوت عن صوت آخر، وبناء نموذج يقيس جودة النطق ويعطي تغذية راجعة عن مكان لفظ الصوت وصفاته. قمنا بدراسة هذه الواصفات على اللغة العربية بشكل أساسي كونها محور البحث، إضافة للّغة الإنجليزية بهدف مقارنة العمل مع الأبحاث المماثلة.

2.4. مخطط عام

تتضمن منهجية العمل الخطوات الموضحة في (الشكل 11- نموذج العمل المقترح)



الشكل 11- نموذج العمل المقترح

اعتمدنا العمل على معطيات موسومة labeled ومقسمة segmented على مستوى الصوتيم. في المرحلة الأولى جرت قراءة البيانات الصوتية ومعالجتها لاستخلاص السمات المناسبة منها، ثم استخدام السمات كدخل لشبكة التعلم، وتتضمن منهجية العمل مرحلتي تعرف، الأولى مهمتها تعرف الصوتيمات، أما الثانية فيتم فيها تعرف واصفات الكلام.

3.4. تحضير المعطيات ومعالجتها

جرى العمل على معطيات من عدة مدوّنات سنتكلم عنها في فقرة **المعطيات المستخدمة** في الفصل التالي. تتم في هذه المرحلة استخلاص معاملات MFCC (Mel Frequency Cepstral Coefficient) بعد أن وجدنا أن معظم الأنظمة تعتمد بشكل أساسي على هذه السمات لأهميتها وقدرتها على تمييز تغيّرات الإشارة الصوتية وفصل أثر المتكلم عن الكلام المنطوق. حيث تتم في البداية قراءة الملفات الصوتية للجمل المنطوقة (المرقمنة بمعدل أخذ عينات 16 كيلو هرتز)، إضافة لقراءة معلومات تقسيم هذه الجمل إلى صوتيماتها (البيانات الموسومة[19])، بعد ذلك يجري تقطيع الإشارة لإطارات زمنية بطول 25 مللي ثانية وانزياح زمني 10 مللي ثانية، ليتم استخلاص السمات الصوتية منها، حيث تعد الإشارة الكلامية مستقرة ضمن النافذة الزمنية السابقة. تحسب معاملات MFCC بالاعتماد على تحويل فورييه السريع (FFT) باستخدام مرشح تمرير حزمة كما في المعادلة (1) [51].

$$\textbf{(1)} \quad \boldsymbol{MFCC(t,k) = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} log\, p_n \cos\left(k(n-0.5)\frac{\pi}{N}\right)}$$

تمثل N عدد مرشحات تمرير الحزمة، p_n تمثل طاقة الخرج من المرشح رقم n في الزمن t، $k = 1,2,,3 \dots L$ حيث L عدد معاملات MFCC.

إضافة إلى معاملات MFCC يتم حساب معامل لوغاريتم الطاقة، ومشتقات هذه المعاملات من الدرجة الأولى والثانية ($\Delta, \Delta\Delta$) لكل إطار صوتي. تضاف بعد ذلك معلومات السياق للإطار الزمني الحالي، إذ بينت الدراسات أهمية إضافة سياق من الإطارات المجاورة للإطار الزمني الحالي لتحسين دقة النموذج [52] (يتم تغذية شبكة التعلم عملياً بعدد إطارات يتراوح بين 9 و 13 إطار زمني تمثل سمات الدخل).

بالنسبة لوحدة تقابل واصفات النطق، يجري فيها تحضير البيانات اللازمة عن طريق ربط كل واصفة مع الصوتيمات التابعة لها في مجموعة المعطيات المدروسة وفق جدول تقابل محدد، يعرض (الجدول 5) توزع الحروف على الواصفات في اللغة العربية حسب [53]، بينما نجد في (الجدول 6) توزعها في اللغة

---

[19] البيانات الموسومة تتضمن الحدود الزمنية لكل صوتيم ضمن الجملة المنطوقة (بدايته ونهايته) مع الوسم label المقابل له، ويمثل الوسم الكتابة الصوتية للصوتيم المقابل للنص المنطوق.

الإنجليزية حسب [47]. نحصل في خرج وحدة التقابل على وسم لكل صوتيم يحدّد إذا كانت الواصفة موجودة فيه أم لا في مجموعة الواصفات التي يتم تدريب النظام عليها.

الجدول 5 واصفات الكلام في اللغة العربية

#	الواصفة Ar	الواصفة En	الحرف
.1	احتكاكي	fricative	س ف خ ذ ز هـ ح ش ظ ث ع غ ص
.2	انفجاري	plosive	ت ك ط ض د ب ق
.3	أمامي	anterior	س ف ت ن ز ذ و ظ ث ل ر ض ط م د ب ص
.4	أنفي	nasal	ن م
.5	بين أسناني	interdental	ظ ذ ث
.6	تكراري	trill	ر
.7	جانبي	lateral	ل
.8	حلقي	pharyngeal	ح ع
.9	حنجري	glottal	هـ ء
.10	حنكي لثوي	alveopalatal	ش ج
.11	شبه صائت	semivowel	و ي
.12	شفوي أسناني	labiodental	ف
.13	شفوي ثنائي	bilabial	و ب م
.14	شفوي طبقي	labiovelar	و
.15	صامت	consonant	س ز ك م د ق ح ظ ث ج ر ت ء هـ ش ل ف ن خ ذ ع غ ض ط ب ص
.16	صمت	silence	صمت
.17	صائت	vowel	ا و ي فتحة ضمة كسرة
.18	طبقي	velar	ك
.19	غاري	palatal	ي
.20	قصير	short	فتحة ضمة كسرة
.21	لثوي أسناني	alveodental	س ت ن ز ل ر ض ط د ص
.22	لهوي	uvular	خ غ ق
.23	مجهور	voiced	ز ي م د ا ظ ج ر و ل ن ذ ع غ ض ب فتحة ضمة كسرة
.24	مدور	rounded	و ضمة
.25	مركب	affricate	ج

.26	مستعل	high	خ ظ غ ط ض ق ص
.27	مستمر	continuant	س ز ي م ا ح ظ ج ث ر و ه ش ل ف ن خ ذ ع غ ض ص فتحة ضمة كسرة
.28	مفخم	emphatic	ط ظ ض ص
.29	مهموس	unvoiced	س ف ت خ ك ح ش ث ط ق ص
.30	هائي	aspirated	ت ك
.31	وسطي	coronal	س ت ن ذ ز ش ظ ث ل ر ض ط د ص فتحة ا

الجدول 6 واصفات الكلام في اللغة الإنجليزية

#	الواصفة Ar	الواصفة En	الحرف
.1	احتكاكي	fricative	ch, dh, f, hh, hv, jh, s, sh, th, v, z, zh
.2	أمامي	anterior	b, bcl, d, dcl, dh, dx, el, em, en, f, l, m, n, nx, p, pcl, r, s, t, tcl, th, v, w, z
.3	أمامي (صائت)	front	ae, aw, ay, eh, ey, ih, iy
.4	أنفي	nasal	em, en, eng, m, n, ng, nx
.5	تقاربي	approximant	r, w, y
.6	جانبي تقاربي	lateral approximant	el, l
.7	حنجري	glottal	hh, hv, q
.8	زمني	tense	aa, ae, ao, ey, ix, iy, ow, oy, uw, ux
.9	شفوي أسناني	labiodental	f, v
.10	شفوي ثنائي	bilabial	b, bcl, em, m, p, pcl
.11	صامت	consonantal	b, bcl, ch, d, dcl, dh, dx, el, em, en, eng, f, g, gcl, hh, hv, jh, k, kcl, l, m, n, ng, nx, p, pcl, q, r, s, sh, t, tcl, th, v, w, y, z, zh
.12	صائت	vowel	aa, ae, ah, ao, aw, ax, ax-h, axr, ay, eh, er, ey, ih, ix, iy, ow, oy, uh, uw, ux
.13	صفيري احتكاكي	sibilant fricative	s, sh, z, zh
.14	صفيري مركب	sibilant affricate	ch, jh
.15	صمت	silence	q
.16	طبقي	velar	eng, g, gcl, k, kcl, ng
.17	غاري	palatal	y

.18	غير صفيري احتكاكي	non sibilant fricative	dh, f, hh, hv, th, v
.19	لثوي	alveolar	d, dcl, dx, el, en, l, n, nx, r, s, t, tcl, z
.20	لثوي حنكي	postalveolar	ch, jh, sh, zh
.21	مجهور	voiced	aa, ae, ah, ao, aw, ax, axr, ay, b, d, dh, dx, eh, el, em, en, eng, er, ey, g, hv, ih, ix, iy, jh, l, m, n, ng, nx, ow, oy, r, uh, uw, ux, v, w, y, z, zh
.22	مدور	round	ao, aw, ow, oy, uh, uw, ux, w
.23	مركزي	central	ax, ax-h, axr, er, ix, ux
.24	مستمر	continuant	aa, ae, ah, ao, aw, ax, ax-h, axr, ay, dh, eh, el, er, ey, f, hh, hv, ih, ix, iy, l, ow, oy, r, s, sh, th, uh, uw, ux, v, w, y, z, zh
.25	مغلق	close	ih, ix, iy, uh, uw, ux
.26	مفتوح	open	aa, ae, aw, ay
.27	وسطي	mid	ah, ao, ax, ax-h, axr, eh, er, ey, ow, oy
.28	وقفي	stop	b, bcl, d, dcl, g, gcl, k, kcl, p, pcl, q, t, tcl

4.4. مرحلة التعرف وتدريب شبكات التعلم

اعتمدت أنظمة تعرف الكلام الآلي ASR لفترة طويلة على الشبكات العصبونية الصنعية، ولم يكن لها تأثير حقيقي في هذا المجال حتى عام 2010، حيث كانت معتمدة على الشبكات العصبونية الضحلة والنماذج التوليدية الاحتمالية HMM, GMM. وقد تجلت صعوبة التحسين لهذه النماذج بسبب ضعف بنية الارتباط الزمني في النماذج التنبؤية العصبونية إضافة لعدم توفر كمية كافية من بيانات التدريب وضعف القدرات الحسابية قبل عام 1990. أدى التوسع في كمية البيانات الصوتية الموسومة وتطور قدرات الحوسبة بالاعتماد على وحدات المعالجة الرسومية GPUs إلى نمذجة بنى متطورة للشبكات العصبونية وهي شبكات التعلم العميق DNNs التي أظهرت تحسناً في دقة تعرف الصوتيمات مقارنة بالنماذج السابقة ضمن التجارب التي تم إجراؤها. ساهم هذا التطور في فتح مجال جديد للباحثين في مجال تعرف الصوت لاعتماد منهجيات التعلم العميق بسبب التحسينات التي تمت على عدة مستويات [7]، منها إمكانية التعامل مع السمات الأولية وعدم الحاجة لتصميم السمات يدوياً، إضافة إلى قدرات التحسين الأفضل كالتغلب على مشاكل فرط التلبيق overfitting، من ناحية أخرى تم الوصول إلى مناعة أعلى ضد الضجيج في تطبيقات تعرف الصوت. فضلاً عن القدرة على نمذجة شبكات التعلم من أجل

لغات مختلفة باستخدام التعلّم المتعدد المهام Multi-task learning[20] والتعلّم المنقول transfer learning. تم ذلك بواسطة البنى الأفضل للشبكات التي جرى بناؤها وتحسينها كشبكات CNNs وشبكات RNNs وبشكل خاص شبكة التعلّم العميق ذات الذاكرة طويلة قصيرة الأمد LSTM Long-short term memory بعد أن حققت معدل خطأ منخفض لتعرف الصوت مقارنة بما سبقها وأظهرت فعالية عالية في التطبيقات الصوتية [54]. تم تصميم هذا النوع من الشبكات لنمذجة التسلسل الزمني والتعامل مع مهام تعرف التسلسل وهذا ما دفعنا لاستخدامها في البحث بهدف تعرف الصوتيمات.

بالنسبة لواصفات الكلام وباعتبارها مسألة تصنيف ثنائي فقد توجهنا لاستخدام شبكات التعلم العميق Deep Neural Network (DNN) بعد أن أثبتت فعاليتها في مسائل التصنيف الثنائي بشكل عام وبصورة خاصة عند تدريب واصفات الكلام في الأعمال السابقة التي تمت دراستها [44] ,[11]. وقد حققت معدل خطأ وسطي أقل بنسبة 56% عبر جميع الواصفات مقارنة بشبكة عصبونية ضحلة بطبقة خفية واحدة [55]. جرى في الدراسات السابقة بناء شبكة تعلم منفصلة لكل واصفة وهذا يستغرق وقتاً أطول وجهداً أكبر، وكون السمات الصوتية مشتركة بين الواصفات توجهنا لتطبيق التعلم المتعدد المهام لتنفيذ تعرف واصفات الكلام لمعرفة أثر تدريب مجموعة واصفات بينها سمات مشتركة على نتائج التعرف.

1.4.4. تعرّف الصوتيمات

تجري عملية تعرف الصوتيمات باستخدام السمات الصوتية التي تم استخلاصها. يمكن القيام بعملية التعرف باستخدام عدة أنواع من شبكات التعلم، وقد اخترنا شبكة LSTM شبكة التعلّم العميق ذات الذاكرة الطويلة القصيرة الأمد Long-short term memory كوننا نتعامل مع بيانات زمنية تسلسلية، فالأداء يصبح أفضل عندما نكون قادرين على الحصول على معلومات سابقة ولاحقة للمعلومة الحالية في مهام التوقع والتعرف. جرى توظيف هذا النوع من الشبكات للتعامل مع البيانات المتسلسلة sequential data في العديد من التطبيقات وحققت نتائج جيدة فيها [57] ,[56].

يتم تحضير السمات بعد استخلاصها لتناسب الشبكة التي سيتم تدريبها، حيث تقسم البيانات إلى ثلاث مجموعات (تدريب وتحقق واختبار)، كما تُقسَّم كل مجموعة لدفعات صغيرة mini batches، تعالج كل دفعة على حدة خلال التدريب. يساعد تدريب الشبكة بشكل دفعات على تحقيق فعالية أعلى للنموذج المدرب والوصول لحل أمثل بشكل أسرع، كما يجعل النموذج قادراً على التعميم بشكل أكبر. إن البيانات الصوتية التي نتعامل معها لها أطوال مختلفة، لذا يتم توحيد طول البيانات لكل تسلسل في كل دفعة وفق

[20] التعلم المتعدد المهام هو أحد منهجيات تعلّم الآلة الذي يهدف لحل العديد من المشكلات المرتبطة معاً، بالاستفادة من التمثيل المشترك بينها.

أطول جملة منطوقة بإضافة حشو صفري للتسلسل الزمني (Zero padding) يتكون من أصفار تضاف في نهاية التسلسل.

بالنسبة لبنية شبكة التدريب يضبط دخل الشبكة حسب عدد السمات المستخدمة في كل مجموعة معطيات. تتكون الشبكة من طبقات خفية عددها n حجم كل طبقة m وهي شبكة ثنائية الاتجاه bidirectional، أما طبقة الخرج فهي طبقة خطية عدد المخارج فيها بعدد الصوتيمات التي يتم تصنيفها في الشبكة.

لتحسين أداء النموذج خلال التدريب يتم استخدام تابع خسارة الإنتروبية التقاطعية Cross Entropy Loss (المعادلة 2) وهو تابع الخسارة الأكثر استخداماً في مسائل التصنيف. يزيد هذا التابع من احتمال تقارب الخرج الحقيقي من خرج الشبكة. يأخذ التابع قيمة بين 0 و1 وكلما كان الخرج أقرب من 0 كان الأداء أفضل. تُستخدم خوارزمية التحسين Adam[21] لتحقيق أقل خسارة ممكنة لتابع الإنتروبية التقاطعية، وتأخذ هذه الخوارزمية معدل تعلّم [22] learning rate يتم ضبط قيمته بالتجريب.

$$\text{Loss} = -\sum_{c=1}^{M} y_{o,c} \log\left(p_{o,c}\right) \qquad (2)$$

حيث:

M عدد الأصناف

log اللوغاريتم الطبيعي

y محدد ثنائي يدل على صحة التصنيف للصنف c ضمن العينة o

p احتمال توقع العينة o للصنف c , $p \neq 0$

[21] خوارزمية التحسين ADAM (Adaptive Moment Estimation) تمتاز بفعاليتها الحسابية والمتطلبات المنخفضة للذاكرة وملائمتها لعدد واسع من مسائل التحسين في مجال تعلم الآلة.

[22] معدل التعلّم هو موسط يستخدم عند تدريب الشبكة العصبونية، يتحكم في مقدار تغيير النموذج المدرب استجابةً للخطأ المقدر في كل مرة يتم فيها تحديث أوزان النموذج، يأخذ قيمة موجبة بين 0 و 1.

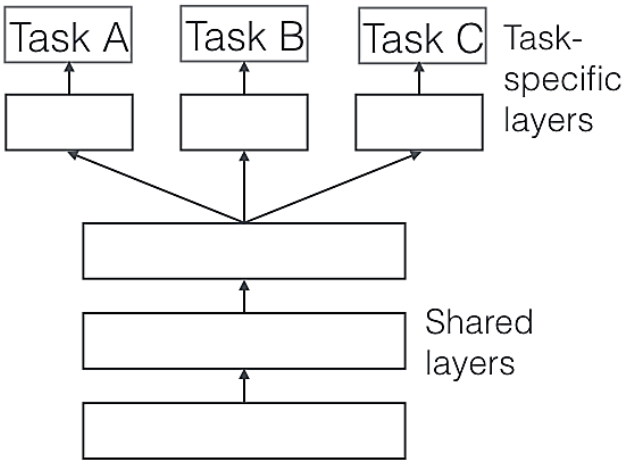
2.4.4. تعرّف واصفات الكلام

تعد واصفات الكلام من السمات المميّزة والهامة كونها ترتبط بعملية نطق الكلام بشكل غني ومفيد، ويعبر عنها بمخارج الحروف وصفاتها. ما يميز هذه السمات أنها تساعد في كشف النطق الخاطئ، حيث يؤدي وجود تغير في أحد هذه الواصفات لحدوث تغير في صوت الحرف وبالتالي إلى خطأ في نطقه. كما أنها مشتركة بين اللغات، وهذا يُمَكّن من الاستفادة من مجموعات معطيات صوتية من لغات مختلفة لتدريب نموذج صوتي لهذه الواصفات. يتم تدريب هذه الواصفات باستخدام مصنفات ثنائية، كون المسألة تهدف لتصنيف الإطارات الصوتية الحاملة للواصفة عن باقي الإطارات، حيث يحتمل كل إطار صوتي حالتين فقط لكل واصفة (الواصفة موجودة أو غير موجودة).

يمكن القيام بهذه المرحلة عن طريق بناء مصنف ثنائي لكل واصفة، لكن لكون السمات مشتركة بين الواصفات توجهنا لاستخدام التعلم المتعدد المهام Multi-task learning [22]. يركز هذا النوع من التعلم على تعميم المعرفة بين المهام المرتبطة، بحيث تضيف كل مهمة المعلومات الخاصة بها وتنقلها لنموذج التعلم لتسهم في تعليم النموذج. في مسألتنا المدروسة نعد تعرف كل واصفة أنه مهمة تصنيف ثنائي، يمكن لهذه المهام أن ترتبط بإحدى طريقتين [58]:

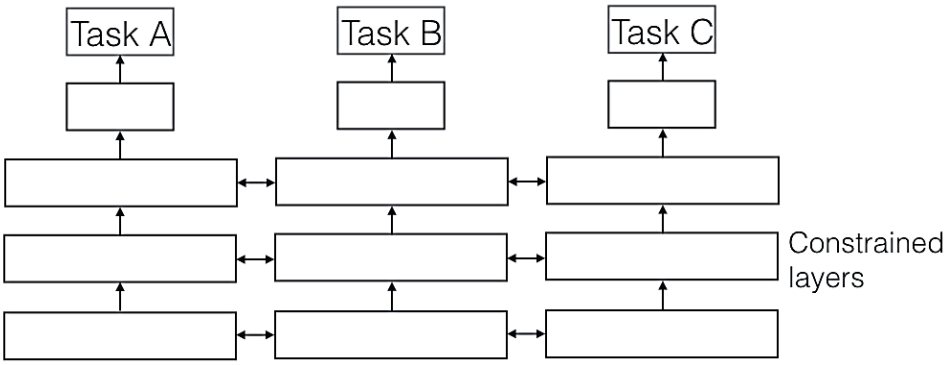
- التشارك القاسي للموسطات hard parameter sharing
- التشارك اللين للموسطات Soft Parameter Sharing

يتم مشاركة الطبقات الخفية في التشارك القاسي للموسطات بين جميع المهام (أي تتشارك المهام بموسطات النموذج وأوزان الطبقات الخفية)، ويكون هناك طبقات خرج منفصلة لكل مهمة (الشكل 12)، وهي المقاربة الأكثر استخداماً في التعلم المتعدد المهام. حيث تتميز هذه الطريقة بأنها تقلل من احتمال حدوث فرط التلبيق overfitting.



الشكل 12– التشارك القاسي للموسطات في التعلم المتعدد المهام

أما في التشارك اللين للموسطات فيكون لكل مهمة نموذج خاص بها مع موسطاته الخاصة، لكن يتم وضع قيود معينة تحدد مدى التشابه بين موسطات نماذج المهام المختلفة (الشكل 13)



الشكل 13– التشارك اللّين للموسطات في التعلم المتعدد المهام

سنعتمد في العمل التشارك القاسي للموسطات hard parameter sharing كونه يعطي أداءً جيداً ويحتاج سعة تخزين أقل على عكس التشارك اللين فهو محدود بقابليته للتوسيع ويميل فيه حجم الشبكة للنمو بشكل خطي مع زيادة عدد المهام.

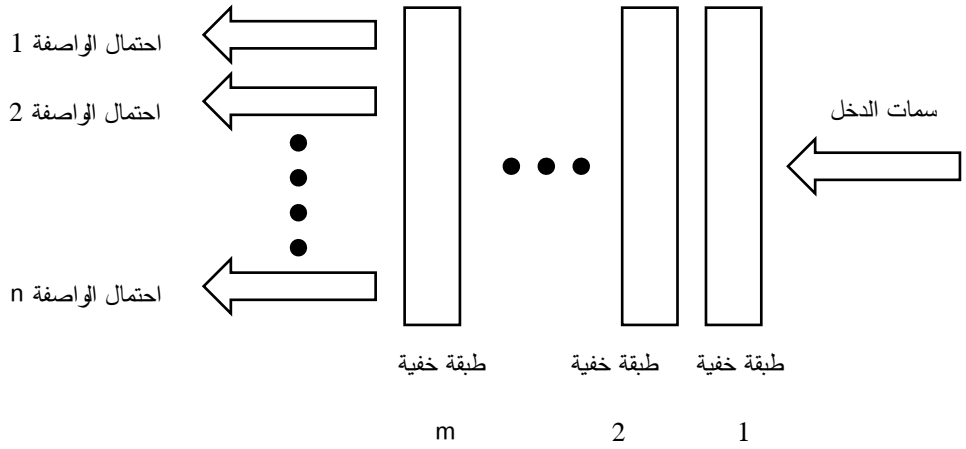
سنعتمد في تعرف واصفات الكلام على نفس السمات المستخلصة من أجل مرحلة تعرف الصوتيمات.

قمنا ببناء شبكة تعلم عميق متعددة المهام للتصنيف الثنائي لهذه الواصفات. تتكون شبكة التعلم من طبقة دخل وعدة طبقات خفية مشتركة بين مهام التصنيف، أما طبقة الخرج فيرتبط حجمها بعدد الواصفات التي يتم تعرفها. يجري تدريب النموذج بحيث يتم تقليل الخطأ الناتج عن كل مهمة وفق خوارزمية التحسين Adam وهي تهدف لتقليل تابع الخسارة الثنائي للإنتروبية التقاطعية binary cross entropy loss (المعادلة 3)، وتكون قيمة الخسارة المطلوب تقليلها هي مجموع قيم الخسارات الناتجة عن تصنيف كل مهمة. يوضح (الشكل 14) البنية العامة في مرحلة تعرف الواصفات.

$$binary\ \text{loss} = -(\text{y} \log(\text{p}) + (1-\text{y}) \log(1-\text{p})) \quad (3)$$

حيث:

y محدد ثنائي (0 أو 1) يدل على صحة التصنيف ، و p احتمال توقع الصنف



الشكل 14– البنية العامة لشبكة تعرف واصفات الكلام

5.4. خاتمة

عرضنا في هذا الفصل منهجية العمل في البحث بدءاً من مرحلة معالجة المعطيات وحتى الوصول لنتيجة التعرف. تضمنت المقاربة المقترحة لكشف خطأ النطق مرحلتين للتعرف، الأولى للتأكد من لفظ الصوتيم والثانية لتحديد نوع خطأ النطق باستخدام واصفات النطق. وقد اعتمدنا في المرحلة الثانية على التعلم المتعدد المهام لمجموعة الواصفات المدروسة كونه يناسب المسألة المطروحة بهدف الوصول لنتيجة تعرف أفضل مقارنة بمقاربات أخرى تعتمد التعلم بشكل فردي لكل واصفة. سنعرض في الفصل التالي التطبيق العملي لهذه المقاربة على مجموعات المعطيات المدروسة.

الفصل الخامس: الاختبارات والنتائج

1.5. المعطيات المستخدمة

جرى العمل على معطيات من عدة مدوّنات، اثنتين منها باللغة العربية وواحدة باللغة الإنجليزية فيما يلي تفصيل كل منها:

1.1.5. مجموعة معطيات النطق بالعربية Arabic Speech Corpus

هي مجموعة معطيات صوتية للغة العربية الفصحى تم بناؤها كجزء من أطروحة دكتوراة نوار حلبي [59] في جامعة ساوثهامتون. وتحوي تسجيلات صوتية بأطوال زمنية مختلفة عددها 1913، مدتها حوالي 4 ساعات لمتحدث عربي الأصل مع النص المقابل لها. جرى تقسيم هذه البيانات على مستوى الصوتيم واستخدامها لتركيب كلام منطوق آلياً ذي جودة عالية. كما يمكن استخدامها في تطبيقات صوتية أخرى. أُتيحت مجموعة المعطيات هذه للاستخدامات غير التجارية عبر موقع خاص بالمجموعة [23] تحت رخصة المشاع الإبداعي [24].

لم يسبق العمل على هذه المعطيات لنفس هدف البحث، لكن جرى توظيفها في أنظمة أخرى، منها نظام لنمذجة المدة الزمنية للكلام [60]، ونظام تحويل نص إلى كلام بالاعتماد على التعلّم العميق [61].

- **محتويات مجموعة المعيطات**

قام الباحثون ببناء هذه المجموعة وفق الخطوات التالية: في البداية تم تجميع النص من موقع الجزيرة لتعليم اللغة العربية[25] إضافة لنصوص ليس لها معنى تم توليدها آلياً للحصول على تشكيلات مختلفة للثنائيات الصوتية diphones، جرى بعد ذلك تسجيل هذه النصوص في استديو خاص من قبل متحدث واحد عربي الأصل بتردد أخذ عينات 48.0 kHz، وبلغ عدد الكلمات في الملفات الصوتية المسجلة حوالي 17040 كلمة، بعد ذلك تم توليد التمثيل الصوتي phonetic transcript آلياً، يحوي هذا التمثيل 82 صوتيماً (الجدول 7)(بالنسبة للصوامت المضعّفة لم يتم ذكرها في الجدول اختصاراً لكن تم تمثيلها برمز مضاعف مماثل لرمز الصوتيم بدون تضعيف مثل الباء المشددة (بّ) تم تمثيلها ب (BB)). بعد التمثيل جرى تقطيع الأصوات بداية على مستوى الكلمة ثم على مستوى الصوتيم وتذييلها بوسم صوتي محدد (اسم الصوتيم) مع إيجاد الحدود الزمنية لبداية ونهاية كل صوتيم حيث جرت عملية التقطيع بشكل آلي، ثم تم القيام بعملية المحاذاة القسرية للصوت forced alignment مع الوسم الصوتي الموافق, بعد ذلك قامت

---

[23] http://en.arabicspeechcorpus.com/

[24] https://creativecommons.org/licenses/by/4.0/

[25] https://learning.aljazeera.net/en

مجموعة من ثلاثة خبراء لغويين بتدقيق العمل للتأكد من محاذاة الصوت زمنياً بشكل صحيح مع النص وتكرار العملية السابقة للحصول على دقة مناسبة.

الجدول 7 مجموعة صوتيمات مجموعة معطيات النطق بالعربية

i1	*[ـِ]*	*u0*	*ـُ*	y	ي	g	غ	r	ر	<	أ
uu1	[و]	*i0*	*ـِ*	v	ڤ	f	ف	z	ز	B	ب
ii1	[ي]	*AA*	*ا*	p	پ	Q	ق	s	س	T	ت
U1	*([ـُ])*	*UU0*	*و*	G	ج	k	ك	$	ش	^	ث
I1	*([ـِ])*	*II0*	*ي*	J	ج (d͡ʒ)	l	ل	S	ص	J	ج (ʒ)
UU1	([و])	*A*	*(ـَ)*	*aa*	*ا*	m	م	D	ض	H	ح
II1	([ي])	*U0*	*(ـُ)*	*uu0*	*و*	n	ن	T	ط	X	خ
sil	pause	*I0*	*(ـِ)*	*ii0*	*ي*	h	ه	Z	ظ	D	د
Dist	distortion	*u1*	*[ـُ]*	*a*	*ـَ*	w	و	E	ع	*	ذ

في الجدول 7 يشير العمود الأيمن في كل قسم للصوتيم بالخط العربي، بينما يشير العمود الأيسر لترميز Buckwalter[26] (تم استخدام نسخة معدلة منه). تشير الرموز التي تحتها خط إلى الحروف التي يتم نطقها في الكلمات الأجنبية مثل فيديو، أما الرموز المائلة فتشير إلى الصوائت، وتشير باقي الرموز إلى الصوامت. تشير الحروف ضمن قوسي "()" لحروف جوارها مفخم، بينما تشير الحروف ضمن قوسي "[]" لحروف ممالة[27] عند اللفظ.

عند العمل على تصنيف الصوتيمات السابقة في هذا البحث جرى تقليص الصوتيمات في مجموعة تضم 38 صوتيماً وذلك بعد دمج الصوت والصوت المشدّد منه، ودمج الحركة وحرف المد التابع لها (مثل الفتحة والألف)، إضافة لاعتبار الأصوات الأجنبية مثل(P,v,G ,..) مع التشويش (dist) صوتيماً واحداً. يوضح الجدول 8 كيفية تقليص الصوتيمات، يمثل العمود الأيمن الصوتيم المقابل بعد التقليص (الصمت لم يتم ذكره باعتبار لا مقابل له)، ويوضح الجدول 9 الصوتيمات بعد التقليص.

---

[26] ترميز Buckwalter تمثيل يستخدم للمقابلة بين الحروف العربية والحروف والرموز اللاتينية.

[27] الإمالة هي النطق بالألف الممالة بين الألف والياء الصحيحتين (تقريب الفتحة من الكسرة، والألف من الياء).

الجدول 8 التقابل بين 82 إلى 38 صوتيم في مجموعة صوتيمات مجموعة معطيات النطق بالعربية

الصوتيم	الصوتيمات المقابلة	الصوتيم	الصوتيمات المقابلة
<	<,<<,AH,Ah	g	g,gg
b	b,bb	f	f,ff
t	t,tt	q	q,qq
^	^,^^	k	k,kk
j	j,jj	l	l,ll
H	H,HH	m	m,mm
x	x,xx	n	n,nn
d	d,dd	h	h,hh
*	*,**	w	w,ww
r	r,rr	y	y,yy
z	z,zz	aa	a,a',aa,aa'
s	s,ss	AA	A,A',AA,AA'
$	$,$$	uu0	u0,u0',uu0,uu0',u
S	S,SS	UU0	U0,U0',UU0,UU0'
D	D,DD	uu1	u1,u1',uu1,uu1',UU1',U1
T	T,TT	ii0	i0,i0',ii0,ii0'
Z	Z,ZZ	II0	I0,I0',II0,II0'
E	E,EE	ii1	i1,i1',I1,I1',ii1,ii1'
		dist	p,pp,v,J,G,dist,-

الجدول 9 صوتيمات مجموعة معطيات النطق بالعربية بعد التقليص إلى 38 صوتيم

الصوتيم	الترميز	الصوتيم	الترميز	الصوتيم	الترميز
ء	>	ض	D	ألف مدية	a
ب	b	ط	T	ألف مدية جوارها مفخم	A
ت	t	ظ	Z	واو مدية	u0
ث	^	ع	E	واو مدية جوارها مفخم	U0
ج	j	غ	g	واو مدية ممالة	u1
ح	H	ف	f	ياء مدية	i0
خ	x	ق	q	ياء مدية جوارها مفخم	I0
د	d	ك	k	ياء مدية ممالة	i1
ذ	*	ل	l	تشويش	dist
ر	r	م	m	صمت	sil
ز	z	ن	n		
س	s	ه	h		
ش	$	و	w		
ص	S	ي	y		

2.1.5. مجموعة المعطيات TIMIT (Texas Instruments Massachusetts Institute of Technology)

مجموعة معطيات صوتية معروفة وشائعة لمتحدثين باللغة الإنجليزية الأمريكية [62]. تم تطويرها من قبل شركة Texas Instruments (TI) ومعهد ماساتشوستس للتكنولوجيا (MIT) ومعهد ستانفورد للأبحاث (SRI)، وهي متوفرة للاستخدام بشكل مجاني[28]. تضم TIMIT تسجيلات لـ 630 متحدث بثماني لهجات مختلفة لكل من الجنسين، قام كل منهم بقراءة عشر جمل، كل جملة بطول 30 ثانية، لتكون المدة الإجمالية لجميع التسجيلات حوالي 5.4 ساعة. تم تقسيم هذه المجموعة إلى بيانات تدريب تضم 462 متحدث وبيانات اختبار تضم بقية المتحدثين البالغ عددهم 168. تحوي المجموعة 45 صوتيماً تم تقليصها في مجموعة تضم 39 صوتيماً كما اقترح [63]. استخدمت TIMIT في العديد من الأبحاث كونها مقطعة وموسومة على مستوى الصوتيم يدوياً. يوضح الجدول 10 الصوتيمات 39 في مجموعة معطيات TIMIT مع توضيح كيفية التقليص.

الجدول 10 صوتيمات مجموعة معطيات TIMIT بعد التقليص إلى 39 صوتيم

رقم الصوتيم	الصوتيم	الصوتيمات المقابلة	رقم الصوتيم	الصوتيم	الصوتيمات المقابلة
1	b	b	21	r	r
2	d	d	22	w	w
3	g	g	23	y	y
4	p	p	24	hh	hh,hv
5	t	t	25	iy	iy
6	k	k	26	ih	ih,ix
7	dx	dx	27	eh	eh
8	jh	jh	28	ey	ey
9	ch	ch	29	ae	ae
10	s	s	30	aa	aa,ao
11	sh	sh,zh	31	aw	aw
12	z	z	32	ay	ay
13	f	f	33	ah	ah,ax,ax-h
14	th	th	34	oy	oy
15	v	v	35	ow	ow
16	dh	dh	36	uh	uh
17	m	m,em	37	uw	uw,ux
18	n	n,en,nx	38	er	er,axr
19	ng	ng,eng	39	sil	pau,epi,h#,bcl,dcl,gcl, pcl,tck,kcl,dcl,tcl,q
20	l	l,el			

[28] https://catalog.ldc.upenn.edu/LDC93S1

3.1.5. مجموعة معطيات الصوتيات العربية KACST Arabic Phonetic Database (KAPD)

هي مجموعة معطيات صوتية تم إنشاؤها من قبل مدينة الملك عبد العزيز للعلوم والتقنية King Abdul-Aziz City for Science and Technology (KACST) [64]. تحوي المجموعة على تسجيلات صوتية لسبعة متحدثين أصليين للّغة العربية لما يقارب 1.2 ساعة، تضم 35 صوتيماً (باعتبار الحرف والحرف المشدد منه صوتيماً واحداً) كما يوضح الجدول 11. قمنا بالعمل في هذا البحث على النسخة المحدثة من هذه المجموعة، حيث طُوِّر جزء منها وقُطِّع يدوياً على مستوى الصوتيم، بهدف استخدامها في تطبيقات تعلّم الآلة والتنقيب عن المعطيات كما يوضح البحث [50]، وقد حصلنا عليها بعد التواصل مع الباحث وطلب استخدام المعطيات لإجراء التجارب عليها. تحوي المجموعة المطورة 4749 تسجيلاً صوتياً لسبعة متحدثين لما يقارب 52 دقيقة. ويُظهر البحث توازن مجموعة المعطيات نسبة لما يقارب 80% من الصوتيمات الموجودة فيها.

الجدول 11 صوتيمات مجموعة المعطيات KAPD

رقم الصوتيم	الصوتيم	رمز الصوتيم	الصوتيمات المقابلة
1	ب	bs10	bs10,bs20
2	ت	ts10	ts10,ts20
3	ث	vs10	vs10,vs20
4	ج	jb10	jb10,jb20
5	ح	hb10	hb10,hb20
6	خ	xs10	xs10,xs20
7	د	ds10	ds10,ds20
8	ذ	vb10	vb10,vb20
9	ر	rs10	rs10,rs20
10	ز	zs10	zs10,zs20
11	س	ss10	ss10,ss20
12	ش	js10	js10,js20
13	ص	sb10	sb10,sb20
14	ض	db10	db10,db20
15	ط	tb10	tb10,tb20
16	ظ	zb10	zb10,zb20
17	ع	cs10	cs10,cs20
18	غ	gs10	gs10,gs20
19	ف	fs10	fs10,fs20
20	ق	qs10	qs10,qs20
21	ك	ks10	ks10,ks20
22	ل	ls10	ls10,ls20,lb20
23	م	ms10	ms10,ms20
24	ن	ns10	ns10,ns20
25	هـ	hs10	hs10,hs20

26	و	ws10	ws10,ws20
27	ي	ys10	ys10,ys20
28	ء	hz10	hz10,hz20
29	َ	as10	as10
30	ا	as20	as20
31	ِ	is10	is10
32	ي	is20	is20
33	ُ	us10	us10
34	و	us20	us20
35	صمت	sil	sil

2.5. مقاييس التقييم

تم إجراء التقييم على مستويين الأول هو مستوى الإطار الزمني frame والثاني هو مستوى الصوتيم phoneme. يجري التقييم على مستوى الصوتيم بالاعتماد على نتيجة تصنيف إطاراته المكونة له حيث يتم اعتماد صنف الإطارات الأغلبية majority. تم اعتماد مقاييس الضبط accuracy، والدقة precision، والاسترجاع recall، والتقييم الشامل f1(F-measure)، ومعدل خطأ الصوتيم Phoneme error rate (PER) عند عمل التجارب.

المقياس	الصيغة
Accuracy	$(TP + TN) / (TN + TP + FN + FP)$
Recall	$(TP) / (FN + TP)$
Precision	$(TP) / (FP + TP)$
F-measure	$(2 \times Precision \times Recall) / (Precision + Recall)$

حيث:

TP موجب حقيقي (True positive) توقع أن الإطار الصوتي يحمل وسماً لصنف محدد مع كون الإطار يحمل وسم الصنف نفسه فعلياً.

TNسلبي حقيقي (True negative) توقع أن الإطار الصوتي لا يحمل وسماً لصنف محدد مع كون الإطار لا يحمل وسم هذا الصنف فعلياً.

FP موجب كاذب (False positive) توقع أن الإطار الصوتي يحمل وسماً لصنف محدد مع كون الإطار لا يحمل وسم هذا الصنف فعلياً.

FN سلبي كاذب (False negative) توقع أن الإطار الصوتي لا يحمل وسماً لصنف محدد مع كون الإطار يحمل وسم هذا الصنف فعلياً.

بالنسبة لمعدل خطأ الصوتيم Phoneme error rate (PER) فيحسب فيه أولاً صنف الصوتيم كما ذكرنا بالاعتماد على صنف الإطارات الأغلبية المكونة له ويكون معدل الخطأ:

معدل خطأ الصوتيم = 1 − نسبة الصوتيمات التي تم تصنيفها بشكل صحيح

3.5. تعرّف الصوتيمات وواصفات الكلام في مجموعة معطيات KAPD

1.3.5. تعرف الصوتيمات في KAPD

تم إجراء عدة تجارب لاختيار الموسطات الأفضل حسب مجموعة معطيات KAPD لصغر حجمها وتنوع عدد المتحدثين فيها مما يعطي ثقة أكبر عند اختبار أداء النموذج المدرب.

أجرينا في البداية عدة اختبارات لمعرفة عدد الإطارات المجاورة التي يمكن إضافة سماتها كدخل للشبكة مع سمات الإطار الأساسي للحصول على نسبة تعرف أفضل.

اعتمدنا التقسيم الأساسي للمجموعة إلى بيانات التدريب والاختبار (71% للتدريب و 29% للاختبار) ويظهر الجدول 13 توزع الصوتيمات في كل منها، تم اختيار 10% من بيانات التدريب لعملية التحقق كما في [49]. في البداية قمنا باستخلاص 8 معاملات MFCC إضافة إلى معامل لوغاريتم الطاقة، ومشتقات المعاملات من الدرجة الأولى والثانية ($\Delta, \Delta\Delta$)، ليصبح طول شعاع السمات 27 للإطار الحالي.

تم ضبط موسطات شبكة LSTM بطبقة خفية واحدة حجمها 1024 لتصنيف صوتيمات هذه المجموعة (35 صوتيم). قمنا بتدريب الشبكة مع تغيير عدد الإطارات الصوتية المجاورة (السابقة واللاحقة) المضاف سماتها لسمات التدريب. ويوضح الجدول 12 نتائج بيانات الاختبار.

الجدول 12 نتائج تعرف الصوتيمات في KAPD باستخدام شبكة LSTM مع عدد إطارات مجاورة مختلفة

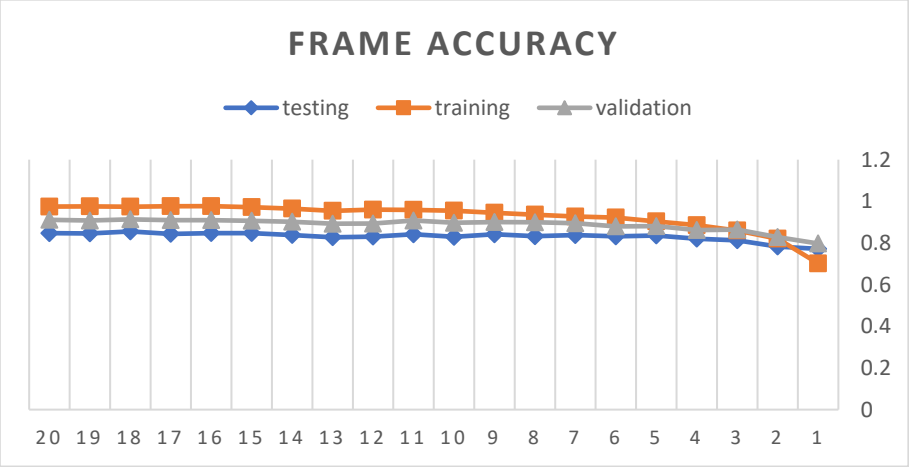
	Phoneme				Frame				
Loss	**recall**	**precision**	**f1**	**Accuracy**	**recall**	**precision**	**f1**	**Accuracy**	**neighbors**
0.53	88.72	88.94	89.54	88.94	83.19	83.51	83.95	83.51	0
0.50	89.54	89.74	90.22	89.74	84.34	84.46	85.32	84.46	1
0.52	89.55	89.87	89.27	89.55	84.34	84.76	84.06	84.34	3
0.49	90.62	90.76	90.99	90.76	85.3	85.54	85.58	85.54	5

يتضح لدينا أن إضافة 5 إطارات صوتية مجاورة يعطي أداء أفضل، لذا سنقوم باعتماده في باقي الاختبارات.

الجدول 13 توزع صوتيمات KAPD في بيانات التدريب والاختبار

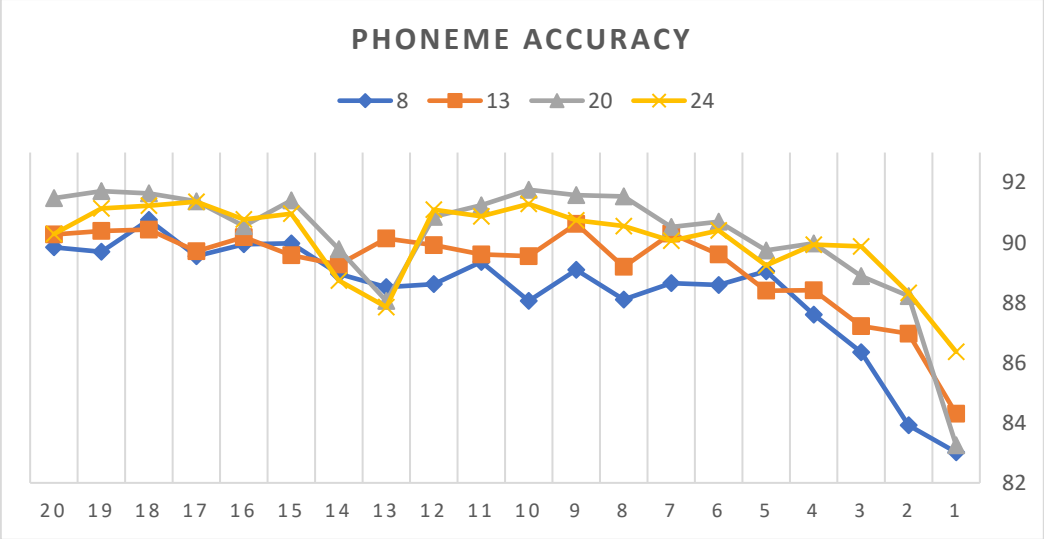
Arabic Phoneme	KAPD Symbol	Training Samples	Testing Samples	Arabic Phoneme	KAPD Symbol	Training Samples	Testing Samples
sil	sil	6767	2715	ع	cs10	121	48
ء	hz10	128	52	غ	gs10	123	48
ب	bs10	121	47	ف	fs10	121	49
ت	ts10	120	49	ق	qs10	120	48
ث	vs10	122	48	ك	ks10	120	48
ج	jb10	117	45	ل	ls10	131	52
ح	hb10	120	48	م	ms10	119	48
خ	xs10	120	48	ن	ns10	119	48
د	ds10	123	51	ه	hs10	131	52
ذ	vb10	121	47	و	ws10	120	47
ر	rs10	120	48	ي	ys10	121	48
ز	zs10	5169	2073	فتحة	as10	1681	671
س	ss10	121	51	ا	as21	30	12
ش	js10	120	48	كسرة	is10	1666	670
ص	sb10	117	48	ي	is21	19	12
ض	db10	106	48	ضمة	us10	1667	664
ط	tb10	119	48	و	us21	20	11
ظ	zb10	123	48	Total		20,283	8138

نعرض في (الشكل 15) أداء الشبكة خلال عملية التدريب بإضافة سمات 5 إطارات سابقة و5 إطارات لاحقة للإطار الحالي باستخدام 8 معاملات MFCC بعد 20 دورة تدريب.



الشكل 15– قياس ضبط الإطار لشبكة LSTM لبيانات التدريب والاختبار والتحقق بدلالة عدد دورات التدريب

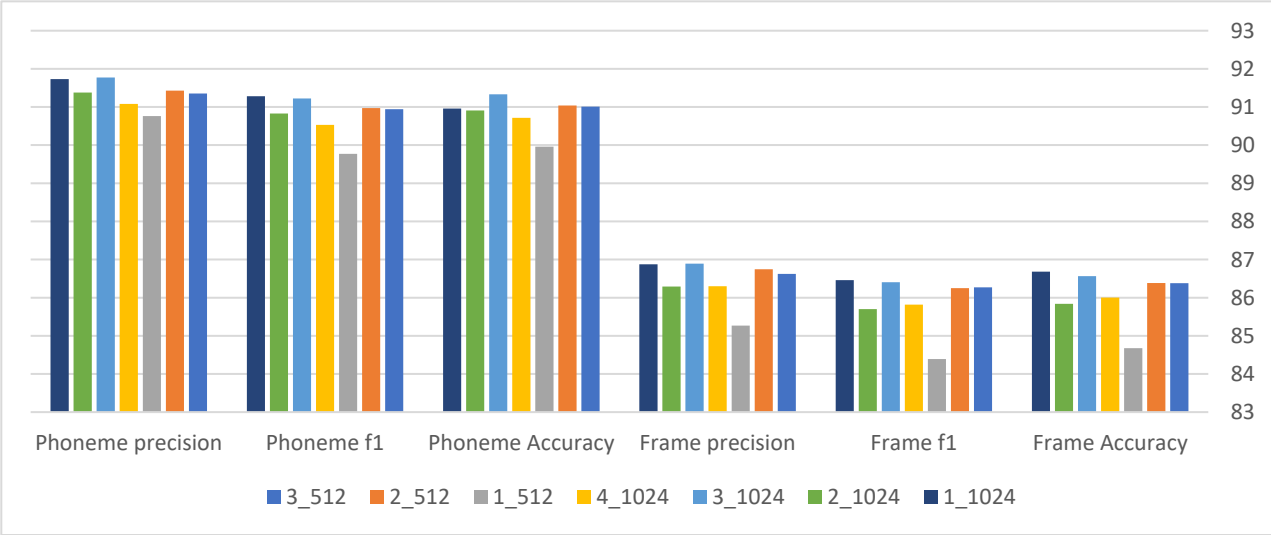
بعد ذلك قمنا بتجربة لتحديد عدد معاملات MFCC، ويوضح (الشكل 16) ضبط التصنيف للصوتيم من أجل قيم مختلفة للمعاملات (8,13,20,24).



الشكل 16– ضبط التصنيف للصوتيم لشبكة LSTM مع تغيير عدد معاملات MFCC بدلالة عدد دورات التدريب

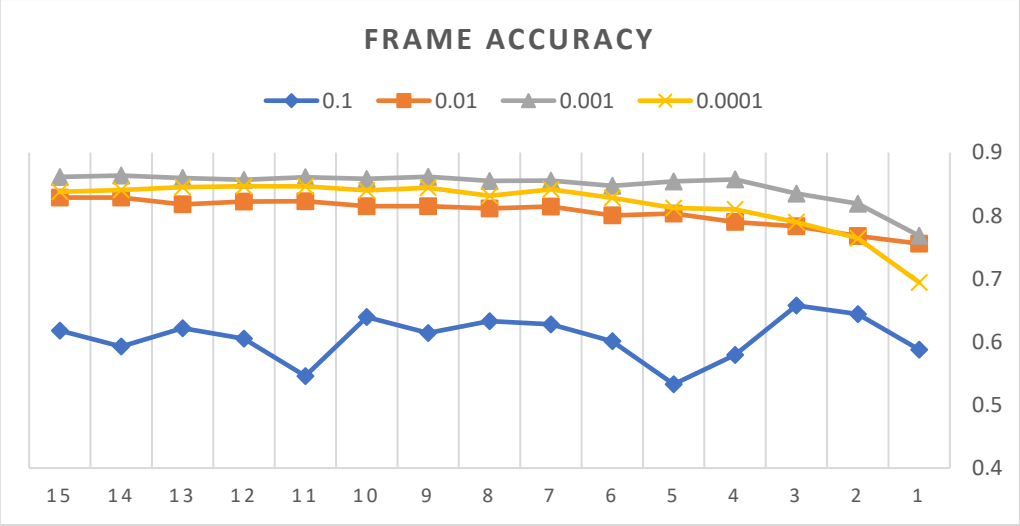
نجد أن استخلاص 20 معامل يعطي أفضل النتائج لذا سنعتمد هذا العدد في باقي التجارب.

قمنا أيضاً باختبار للتحقق من بنية الشبكة المناسبة (عدد الطبقات الخفية وحجمها)، وجدنا تقارب النتائج في التي التجارب التي قمنا بها، مع الحصول على نتيجة جيدة باستخدام طبقة خفية واحدة بحجم 1024 عصبون كما يوضح الشكل 17.



الشكل 17– مقارنة مقاييس التقييم مع تغيير بنية شبكة التعلم (عدد الطبقات وحجم كل منها)

كما جرى اختبار الأداء لتحديد أفضل قيمة لمعامل التعلم وقد حصلنا على أفضل النتائج من أجل قيمة 0.001. يبين (الشكل 18) مقارنة ضبط التصنيف على مستوى الإطار الصوتي من أجل قيم مختلفة لمعامل التعلم



الشكل 18– ضبط التصنيف بدلالة عدد دورات التدريب من أجل قيم مختلفة لمعامل التعلم

يوضح الجدول 14 مقاييس التقييم لكل من الإطار والصوتيم لبيانات الاختبار في مجموعة معطيات KAPD بعد استخلاص 20 معامل MFCC بإضافة سمات 5 إطارات سابقة ولاحقة وفق شبكة تحوي طبقة خفية واحدة حجمها 1024 عصبون.

الجدول 14 مقاييس التقييم لتعرّف الصوتيمات في KAPD باستخدام شبكة LSTM

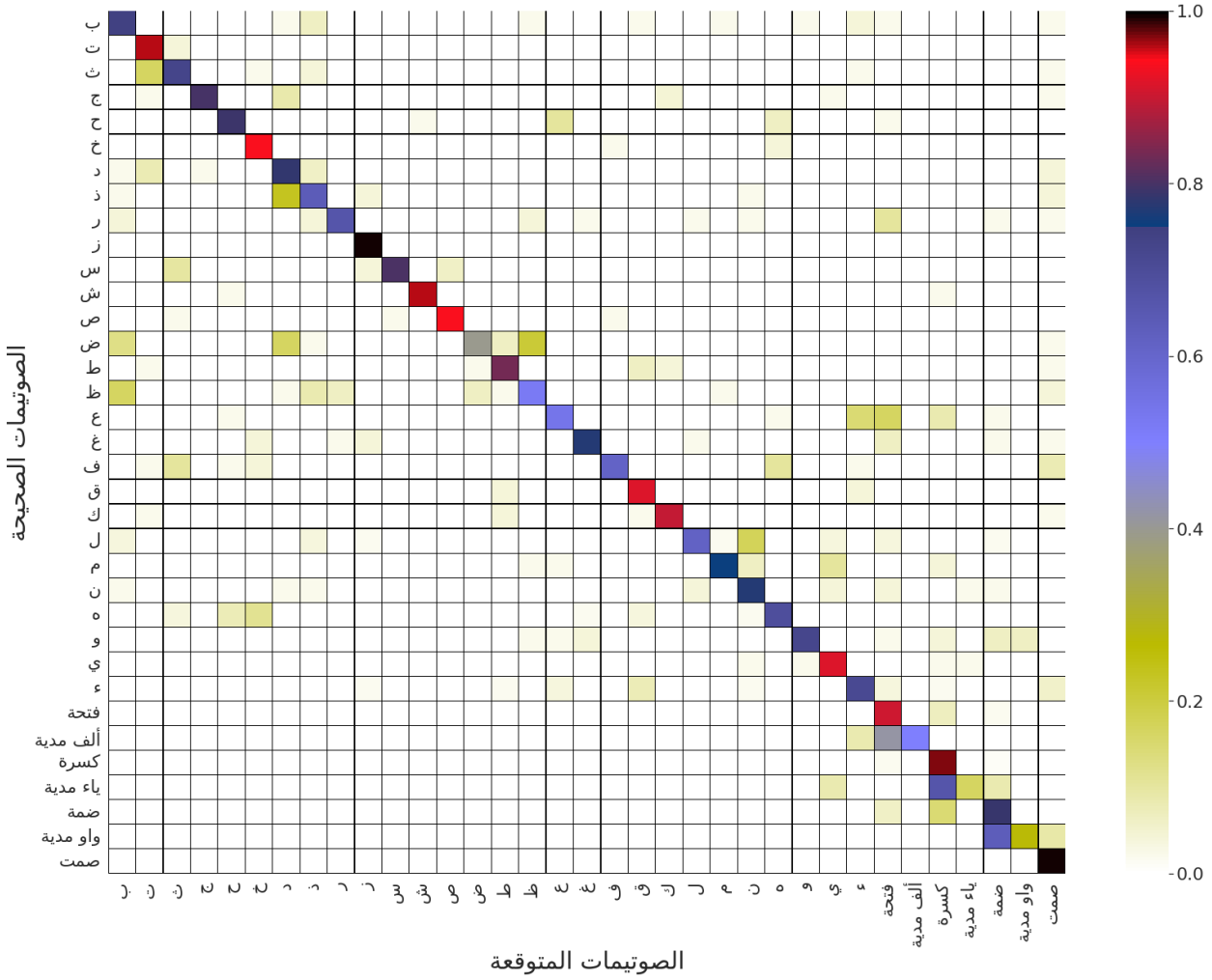
Phoneme	Frame	
91.48	86.68	**Accuracy**
91.73	86.87	**Precision**
91.28	86.46	**F1**

يبين الجدول 15 معدل خطأ الصوتيم PER في KAPD مقارنة مع (Algabri et al., 2021)

الجدول 15 معدل خطأ الصوتيم في KAPD بالمقارنة مع منهجيات أخرى

PER (%)	المنهجية
10.84	AFD–Obj (YOLOv3–tiny–1S)
5.63	PD–Obj (YOLOv3–tiny–2S)
8.52	LSTM (هذا العمل)

إن النظام PD-Obj (phoneme detection object) الذي يستخدم تقنية كشف الأغراض بالاعتماد على الصور الطيفية للإشارة الصوتية YOLOv3 لتعرف الصوتيمات حقق معدل خطأ، بينما تفوقت شبكة LSTM على نظام AFD-Obj (articulatory features detection object) المعتمد على تعرف واصفات النطق في تعرف الصوتيمات باستخدام خوارزمية YOLO لكشف الأغراض أيضاً. يمكن أن نفسر سبب تفوق نظام PD-obj أن الصور الطيفية (Mel-spectrogram) مع مشتقاتها (delta , delta-delta) قادرة على إبراز تغيرات الإشارة وفي حالتنا (إشارة الكلام) تبدو التغيرات واضحة مما يساعد في تحسين التعرف، إضافة لفعالية خوارزمية YOLO. أما بالنسبة لنظام AFD-Obj فهو يقوم بتحديد الصوتيم وفق الواصفات التي تم تعرفها مقارنة مع شعاع واصفات مرجعي لكل صوتيم بنسبة تشابه 100%، وبالتالي فإن خطأ تعرف الواصفات سيؤثر بشكل مباشر على تعرف الصوتيمات.



الشكل 19– مصفوفة الالتباس في مجموعة معطيات KAPD

وجدنا بعد تحليل مصفوفة الالتباس (الشكل 19) أن نسبة التعرف الأقل كانت للصوتيمات التي لها عدد عينات أقل في كل من بيانات التدريب والاختبار مثل (حروف المد الألف والواو والياء)، بالإضافة لوجود لبس في التعرف بين حرف المد والحركة المقابلة له. نلاحظ أيضاً وجود لبس في تعرف صوتيم الضاد مع الصوتيمات (ب، د، ظ) لوجود صفات نطق مشتركة بينها. بالمثل يوجد تداخل في التعرف بين صوتيمي اللام والنون لاشتراكهما في المخرج ومعظم صفات الحروف.

2.3.5. تعرف واصفات الكلام في KAPD

استُخدمت نفس الواصفات المستخدمة في العمل [53] وهي 31 واصفة (راجع الجدول 5 في الفصل الرابع)، وتم استخدام سمات مرحلة تعرّف الصوتيمات نفسها (693 سمة) مع وسم الواصفة المقابل لكل إطار صوتيم. بعد ذلك غُذّيت شبكة تعلم عميق متعددة المهام ذات مصنف ثنائي بهذه السمات. تتكون شبكة التعلم المدرّبة من ثلاث طبقات خفية بعدد مخارج 31، يمثل كل مخرج سمة مختلفة. تم اختيار موسطات شبكة التعلم بمساعدة إطار العمل التحسيني البرمجي Optuna[29] بإجراء عدة تجارب مع تغيير حجم كل طبقة خفية وتابع تفعيل الطبقة، إضافة لتغيير قيمة معدل التعلّم learning rate. ويوضح (الجدول 16) مجال قيم التجارب التي تم إجراؤها.

الجدول 16 مجال قيم التجريب لموسطات شبكة DNN

	حجم الطبقة	تابع التفعيل
الطبقة الخفية الأولى	[250, 2048]	["relu", "sigmoid"]
الطبقة الخفية الثانية	[250, 2048]	["relu", "sigmoid"]
الطبقة الخفية الثالثة	[250, 2048]	["relu", "sigmoid"]
معدل التعلّم	[1e-5, 1e-1]	

أعطت الموسطات التالية أقل نسبة ضياع بالنسبة لبيانات الاختبار.

	حجم الطبقة	تابع التفعيل
الطبقة الخفية الأولى	1321	relu
الطبقة الخفية الثانية	1960	sigmoid
الطبقة الخفية الثالثة	994	relu
معدل التعلّم	1.3657e-05	

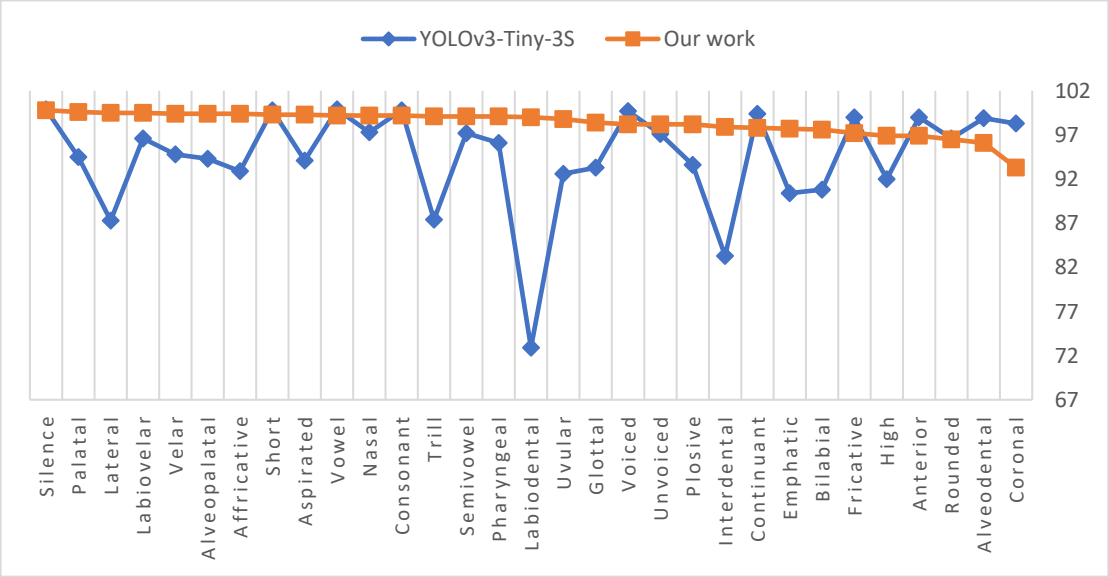
[29] إطار العمل التحسيني البرمجي Optuna هو إطار عمل مفتوح المصدر يساعد على أتمتة عملية إيجاد القيم المثلى للموسطات https://optuna.org/

يبين الجدول 17 و(الشكل 20) نتائج التصنيف على مستوى الصوتيم باستخدام مقياس F1 بالمقارنة مع [49]. تم استخدام مقياس F1 كون البيانات غير متوازنة بالنسبة للواصفات، ونلاحظ تفوق النموذج المدرب بالاعتماد على المصنفات الثنائية بفارق وسطي 4%~، وقد حصلنا على نتائج أفضل بالنسبة ل 66% من الواصفات.

الجدول 17 نتائج مقياس F1 لشبكة DNN لتصنيف واصفات الكلام في KAPD

Our work	YOLOv3-Tiny-3S	
99.4	92.9	Affricative
96.1	98.9	Alveodental
99.4	94.3	Alveopalatal
96.9	99	Anterior
99.3	94.1	Aspirated
97.6	90.8	Bilabial
99.2	99.8	Consonant
97.8	99.4	Continuant
93.3	98.3	Coronal
97.7	90.4	Emphatic
97.2	99	Fricative
98.4	93.3	Glottal
96.9	92	High
97.9	83.3	Interdental
99	72.9	Labiodental
99.5	96.6	Labiovelar
99.5	87.3	Lateral
99.2	97.3	Nasal
99.6	94.5	Palatal
99.1	96.1	Pharyngeal
98.2	93.6	Plosive
96.5	96.6	Rounded
99.1	97.2	Semivowel
99.3	99.8	Short
99.8	99.9	Silence
99.1	87.4	Trill
98.2	97.1	Unvoiced
98.8	92.6	Uvular
99.4	94.8	Velar
98.2	99.7	Voiced
99.2	99.9	Vowel
98.4	94.5	**Average**

يعود السبب في كون نتيجة التعرف أفضل في هذا العمل اعتمادنا على المعطيات الموسومة مسبقاً في تحديد وسم الإطار الصوتي للواصفات. بينما يتم في نظام AFD-Obj تحديد الوسم بناء على الحدود الزمنية للصوتيمات المستنتجة من خرج خوارزمية YOLO، ما يقلل من دقة صحة الوسم للإطار الصوتي في عملية التدريب. كما أن نظامنا يتفوق على نظام [53] المعتمد على نماذج البيرسيبترون متعدد الطبقات MLP وشبكات التعلم العميق، فقد حصل النظام السابق على نسبة أقل من 80% لمقياس f1 لما يقارب 61% من الواصفات لأفضل النماذج المدربة بينما تجاوزت قيمة هذا المقياس 90% لجميع الواصفات في نموذجنا. من جهة أخرى يقوم نظام [53] ببناء نموذج منفصل لكل واصفة ببنى مختلفة وهذا مكلف وغير فعال خاصة مع زيادة عدد الواصفات التي يتم دراستها، إضافة إلى أن النظام السابق يعتمد عدداً ثابتاً من الإطارات الزمنية لكل صوتيم (15 إطار) وهذا يفسر عدم دقة النتائج التي تم الحصول عليها، لأن الصوتيمات مختلفة الطول وتثبيت عدد الإطارات لا يعطي ديناميكية للنموذج بسبب تغير الصوت من متحدث لآخر وتغير الفترة الزمنية للفظ كل صوتيم.



الشكل 20 – نتائج تصنيف واصفات الكلام على مستوى الصوتيم باستخدام مقياس F1

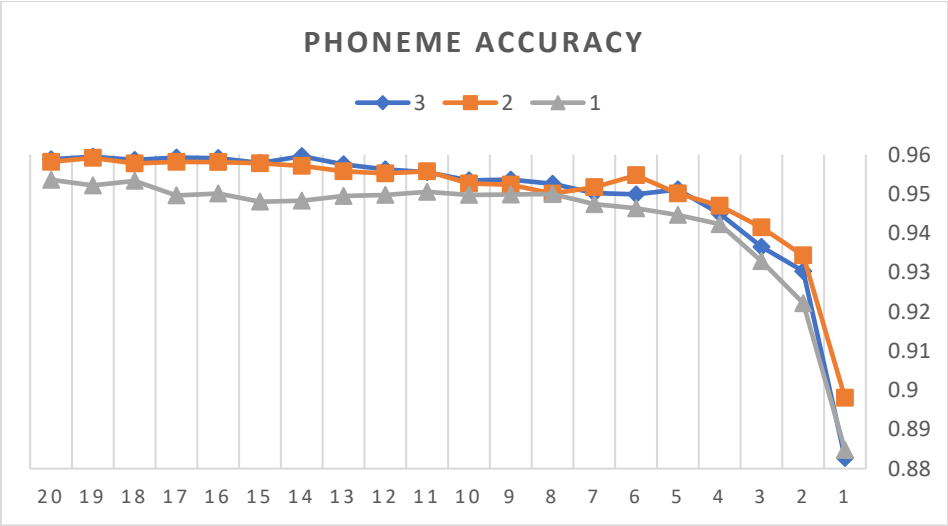
4.5. تعرّف الصوتيمات وواصفات الكلام في مجموعة معطيات النطق بالعربية MSA

1.4.5. تعرف الصوتيمات في مجموعة معطيات النطق بالعربية

تم استخلاص السمات الصوتية بالاعتماد على الموسطات التي أعطت أفضل النتائج مع مجموعة معطيات KAPD (20 معامل MFCC) إضافة إلى معامل لوغاريتم الطاقة، ومشتقات هذه المعاملات من الدرجة الأولى والثانية (Δ, ΔΔ)، ليصبح لدينا شعاع سمات بطول 63 لكل إطار صوتي، وبعد إضافة 11 إطار زمني كمعلومات للسياق يصبح طول شعاع السمات 693 سمة لكل إطار زمني.

تم تقسيم البيانات بنسبة تدريب training 60% ، تحقق Validation 20% ، اختبار testing 20%.

قمنا بتجربة عدة بنى لشبكة LSTM بتغيير عدد الطبقات (1،2،3) مع ضبط حجم كل طبقة (1024 عصبون) وحصلنا على أفضل النتائج نسبياً باستخدام ثلاث طبقات خفية كما يوضح (الشكل 21).



الشكل 21– ضبط التصنيف للصوتيم في مجموعة معطيات النطق بالعربية باستخدام عدد طبقات مختلفة لشبكة LSTM

يبين الجدول 18 نتائج بيانات الاختبار بعد تدريب الشبكة باستخدام 3 طبقات خفية حجم كل منها 1024 عصبون وذلك بعد مرور 20 دورة تدريب epoch.

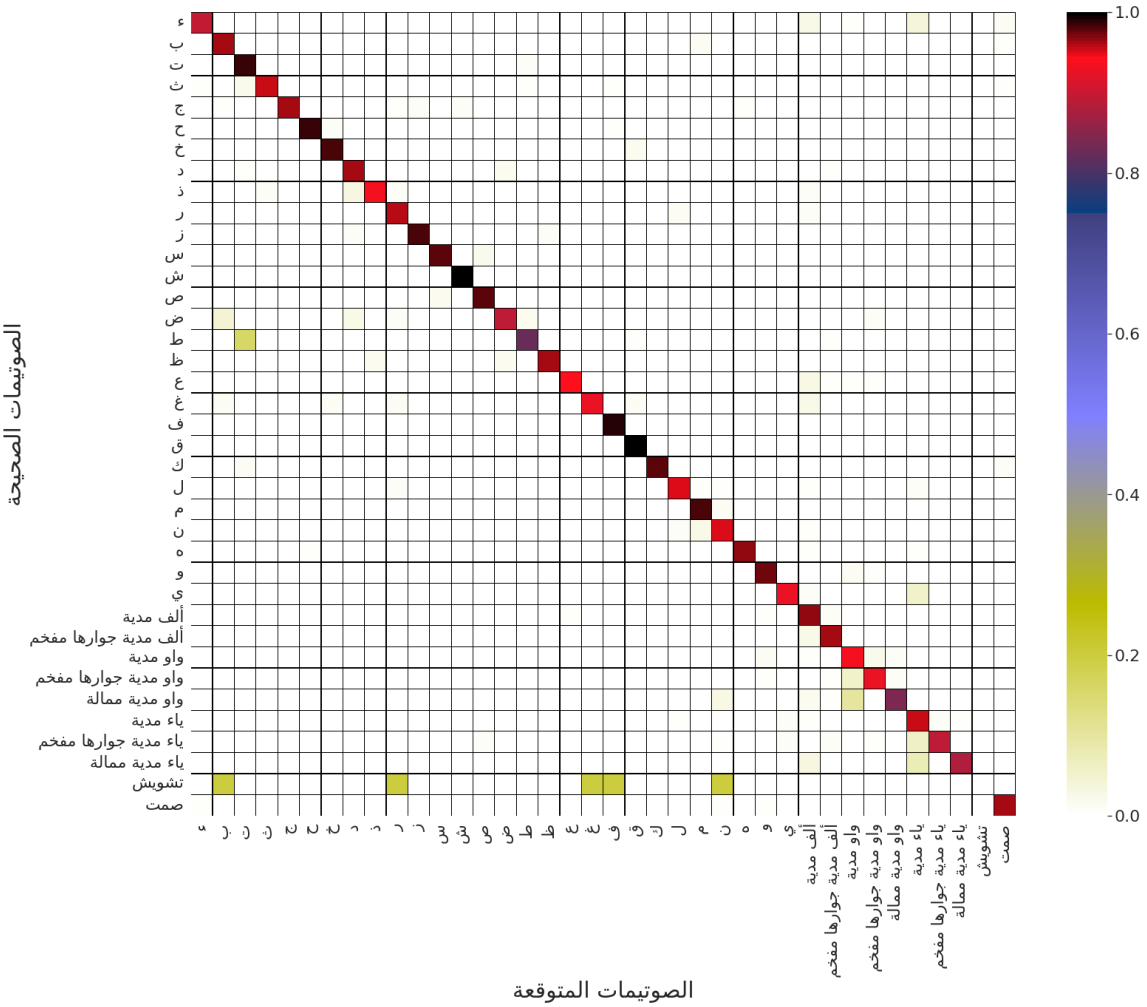
الجدول 18 نتائج تقييم بيانات اختبار مجموعة معطيات النطق بالعربية لشبكة LSTM بثلاث طبقات خفية

F1	Precision	Accuracy	
91.95	91.95	91.97	Frame
95.87	95.87	95.88	Phoneme

لاحظنا أن معظم الأخطاء في التعرّف كانت في الصوتيمات التي تتصف بصفة الشدة (مثل الهمزة)، حيث يكون زمنها قصيراً مقارنة بباقي الصوتيمات. إضافة لوجود لبس في التعرّف بين الصوتيمات المتشابهة مثل (ط–ت) (ض–د)، كون هذه الأصوات تشترك بالمخرج وبعض صفات الحروف. فمثلاً الطاء والتاء تخرجان من طرف اللسان وتشتركان بصفة الشدة، إلا أن الطاء تتميز عنها بوجود الاستعلاء والإطباق فيها، بينما تتصف التاء بالهمس. بالنسبة لحروف المد لاحظنا فيها وجود خطأ تعرّف بين صوت حرف المد وصوت حرف المد الذي يجاوره حرف مفخم لتشابه النطق فيهما. تبين أيضاً وجود لبس في التعرف بين الياء المدية وغير المدية وبعد العودة للبيانات واختبارها اتضح لدينا عدم وجود دقة في وسم هذين الصوتيمين في حالة الياء المشددة (انظر الجدول 19). حيث جرى وسم الياء المشددة بياء مدية (ii0) ملحقة بياء غير مدية (y)، لكن الوسم الصحيح هو ياء غير مدية مشددة مسبوقة بكسرة. أما بالنسبة للتشويش (dist) فكانت نسبة التعرّف فيه منخفضة لعدم وجود عدد عينات كاف منه في بيانات التدريب والاختبار. تتضح هذه الأخطاء في (الشكل 22) لمصفوفة الالتباس confusion matrix للصوتيمات.

الجدول 19– مثال على وسم كلمات تحوي ياء مشددة في مجموعة معطيات النطق بالعربية

الكلمة	الترميز الصوتي
أكاديميّة	a k aa d ii0 m **ii0' y** a t i0 >
فنيٌّ	f a nn **ii0' y** u0
تقليديّ	t A q l ii0 d **ii0' y** i0



الشكل 22– مصفوفة الالتباس لصوتيمات مجموعة معطيات اللغة العربية

2.4.5. تعرف واصفات الكلام في مجموعة معطيات النطق بالعربية

قمنا ببناء نموذج تعرّف على الواصفات في الجدول 2 و 3 (الفصل الثاني) باستخدام مجموعة معطيات اللغة العربية بشكل مماثل لمجموعة معطيات KAPD. تم تقسيم البيانات إلى 40% تدريب، 30% اختبار، 30% تحقق. ويبين الجدول 20 موسطات الشبكة المدربة التي حصلنا عليها عن طريق optuna بعد إجراء عدة تجارب واختيار الموسطات التي أعطت أفضل أداء للتعرف.

الجدول 20 موسطات شبكةDNN لتعرّف واصفات الكلام في MSA

	عدد الطبقات الخفية		3
	رقم الطبقة	عدد الوحدات في الطبقة	تابع التفعيل
خصائص الطبقات الخفية	1	1920	Relu
	2	1949	Sigmoid
	3	498	Relu
تابع الأمثلة	Adam		
Optimizer	Learning rate 6.87e-05		

يعرض الجدول 21 نتائج التصنيف التي حصلنا عليها حيث بلغت دقة الضبط الوسطية لبيانات الاختبار على مستوى الصوتيم 99.17 وهي نسبة تعرّف عالية تسمح باستخدام خرج المصنف كدخل لمرحلة تالية يتم فيها كشف خطأ النطق بالاعتماد على هذه الواصفات. نجد أقل نسبة تعرف في واصفات الحروف الصوتية Vowels وحروف الرخاوة Softness لتنوع الأصوات التابعة لها مما يزيد من معدل الخطأ فيها، ويجعلنا بحاجة لطريقة فعالة أكثر في التمييز كاستخدام مرحلة سابقة تكشف عن واصفات أكثر تحديداً، وتضيق مجال التعرف بالنسبة لهذه الأصوات. نجد أيضاً نسبة تعرف عالية بالنسبة لكل من (Spreading، Affricates، Prolongation، Post alveolar) كونها صفات مميّزة عند النطق بها، عدد أصواتها محدد وهذا يسهل تمييزها في الإشارة الصوتية أيضاً.

الجدول 21 نتائج تصنيف شبكةDNN لواصفات الكلام في مجموعة معطيات اللغة العربية

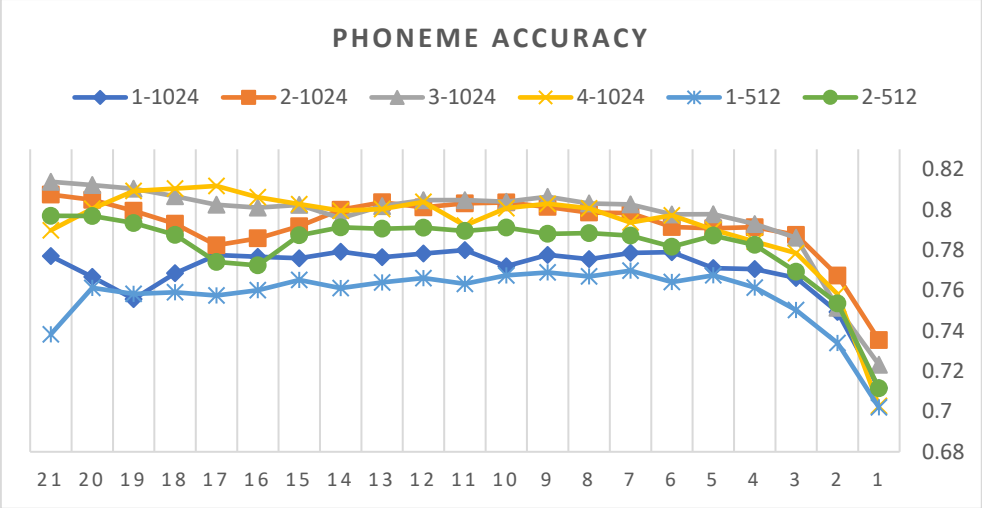
f1	precision	accuracy	
98.77	98.77	98.78	strength
99.36	99.36	99.36	whisper
99.07	99.07	99.08	elevation
97.71	97.73	97.71	softness
96.53	96.66	96.52	vowels
98.93	98.93	98.93	labial
98.39	98.41	98.41	moderate
99.94	99.94	99.94	whistle
99.98	99.98	99.98	spreading
99.16	99.15	99.18	adhesion
98.53	98.59	98.59	pharynx
99.87	99.87	99.87	deep_tongue
98.69	98.80	98.80	middle_tongue
98.48	98.49	98.49	tongue_tip
99.26	99.27	99.28	tongue_border
99.76	99.78	99.79	prolongation
99.41	99.42	99.42	repetition
99.88	99.88	99.88	hiding
99.07	99.07	99.07	fricatives
99.94	99.94	99.94	affricates
98.92	98.92	98.93	silence
99.81	99.81	99.81	interdental
99.25	99.27	99.28	glottal
99.46	99.47	99.48	pharyngeal
99.90	99.90	99.89	uvular
99.21	99.22	99.22	deviate
99.96	99.96	99.96	post-alveolar

بالمقارنة بين الواصفات المدروسة في كل من مجموعة معطيات النطق بالعربية ومجموعة معطيات KAPD نجد أن الواصفات المختارة في KAPD متنوعة أكثر وتركز في بعض الصفات على الحروف متشابهة المخرج مثل (ظ ذ ث) التي تخرج من بين الأسنان، فيمكن تمييز الظاء فيها عن طريق صفة الاستعلاء، بينما تميز الثاء من خلال صفة الهمس فيها، وهذا يمكن من تمييز خطأ النطق فيها بدقة. على سبيل المثال أيضاً صوتيما الصاد والسين يشتركان في المخرج ويتشابهان في بعض الصفات إلا أن الصاد تتميز عن السين بوجود صفة الاستعلاء والتفخيم فيها. بالنسبة لصوتيم الضاد والذي وجدنا فيه لبس في التعرف مع عدة صوتيمات نحتاج صفة إضافية مميزة له وهي الاستطالة (جرى دراستها مع واصفات النطق بالعربية). إن هذا التحليل للصوتيمات وواصفاتها يساعد في انتقاء الواصفات المناسبة لنكون قادرين على كشف خطأ النطق بدقة وهذا ما سنعمل عليه في الأبحاث القادمة.

5.5. تعرّف الصوتيمات وواصفات الكلام في مجموعة معطيات TIMIT

1.5.5. تعرف الصوتيمات في مجموعة معطيات TIMIT

تم تدريب شبكة تعلّم عميق لتصنيف الصوتيمات 39 باعتماد مجموعة اختبار من 24 متحدث core test set من أجل عملية الاختبار إضافة لمجموعة الاختبار الكاملة المكونة من 168 متحدث. قمنا باستثناء الجمل SA (جمل اللهجة) من مجموعة المعطيات قبل التدريب كما هو موصى به في [65] كونها تمثل الجمل التي تم نطقها من قبل المتحدثين للكشف عن تغيرات اللهجة. تم استخلاص السمات الصوتية (13 معامل MFCC)، ومشتقاتها من الدرجة الأولى والثانية ($\Delta, \Delta\Delta$)، ليصبح لدينا شعاع سمات بطول 39 لكل إطار صوتي (باعتبار أن هذا العدد من المعاملات تم استخدامه مع مجموعة معطيات TIMIT في العديد من الأبحاث التي تم الاطلاع عليها). وبعد إضافة سمات 5 إطارات مجاورة سابقة ولاحقة يصبح طول شعاع السمات 429. تم تجريب عدة بنى لشبكة التدريب بتغيير عدد الطبقات الخفية وحجم كل منها ويوضح (الشكل 23) نتائج ضبط التصنيف لمجموعة الاختبار.



الشكل 23– ضبط التصنيف في مجموعة معطيات TIMIT من أجل بنى مختلفة لشبكة LSTM

حصلنا على أفضل أداء من أجل ثلاث طبقات خفية بحجم 1024، ويوضح **الجدول 22** نتائج تدريب هذه الشبكة لكل من مجموعتي الاختبار الكاملة والجزئية بعد 20 دورة تدريب.

الجدول 22 مقاييس التقييم لتعرّف الصوتيمات في TIMIT باستخدام شبكة LSTM

Core Testing		Testing		
phoneme	frame	phoneme	frame	
80.71	79	81.38	79.4	Accuracy
80.87	78.94	81.52	79.25	Precision
80.63	78.91	81.25	79.26	F1

يبين الجدول 23 معدل خطأ الصوتيم PER لتعرّف صوتيمات TIMIT لمجموعة الاختبار مقارنة مع بعض الأعمال.

الجدول 23 مقارنة خطأ تعرّف الصوتيم في TIMIT مع نماذج مختلفة

Phoneme error rate (%)	العمل والمرجع
15.89	CenterNet-DLA [48]
13.8	PYTORCH-KALDI [66]
20.36	CNN [67]
22.39	HMM [68]
39	Cascade CNN [[69]
18.62	LSTM (هذا العمل)

نجد بالمقارنة أن أداء شبكة LSTM عند تعرّف الصوتيمات في TIMIT أفضل من كل من HMM و CNN وCascade CNN. وقد اعتمد العمل الأخير على التصنيف لمجموعات جزئية وفق مرحلتين، الأولى يتم فيها تصنيف الصوتيمات لصوامت وصوائت ثم يجري تصنيف كل مجموعة لمجموعات ثانوية جزئية يحوي كل منها صوتيمات معينة. نلاحظ حصول العمل في PYTORCH-KALDI على أفضل أداء وذلك باستخدام مجموعة مختلفة من الشبكات مثل Li-GRU[30] و MLP، واستخدام مجموعة مختلفة من السمات مثل MFCC و FBANK(filter-bank) و fMLLR[31]، كما حقق العمل باستخدام شبكة CenterNet نتيجة جيدة بالاعتماد على تقنية كشف الأغراض.

[30] Li-GRU: Light gated recurrent units.
[31] FMLLR: Feature space maximum likelihood linear regression.

2.5.5. تعرف واصفات الكلام في مجموعة معطيات TIMIT

قمنا باعتماد واصفات الكلام للغة الإنجليزية المستخدمة في العمل [47] وهي 28 واصفة (انظر الجدول 6 في الفصل الرابع) ، وتم استخدام سمات مرحلة تعرّف الصوتيمات (462 سمة) مع وسم الواصفة المقابل لكل إطار صوتيم. بعد ذلك غذّيت شبكة تعلم عميق ذات مصنف ثنائي بهذه السمات. تتكون شبكة التعلم المدرّبة من ثلاث طبقات خفية بعدد مخارج 28، يمثل كل مخرج سمة مختلفة. ويبين الجدول 24 نتائج ضبط التصنيف Accuracy على مستوى الإطار الصوتي مقارنة مع [49] ,[47]، حيث نجد تقارب النتائج الوسطية لهذه الواصفات مع تفوق نموذجنا بنسبة 0.47% و 0.1% على الترتيب، وملاحظة النتائج الأفضل لبعض الواصفات في نموذجنا المدرب (Silence,Voiced ,Continuant)، إضافة لدقة تصنيف عالية بالنسبة للصمت Silence في هذا العمل.

الجدول 24 نتائج ضبط تصنيف شبكة DNN لواصفات الكلام في TIMIT

LAS-MTL	YOLOv3-Tiny-2S	Our Work	
95	91.05	90.47	Alveolar
90	89.69	89.09	Anterior
98	97.12	96.85	Approximant
98	97.70	**98.02**	Bilabial
99	93.73	97.52	Central
97	94.13	94.18	Close
88	88.97	**89.86**	Consonantal
89	91.37	**94.07**	Continuant
95	96.03	**96.28**	Fricative
95	93.33	91.66	Front
99	98.67	**99.37**	Glottal
99	98.88	98.54	Labiodental
99	98.21	97.64	Lateral approximant
97	90.28	90.13	Mid
99	97.59	97.63	Nasal
97	97.60	97.31	Non sibilant fricative
98	96.09	94.91	Open
99	99.60	99.56	Palatal
99	99.18	98.79	Postalveolar
98	94.99	94.73	Round
99	99.5	99.37	Sibilant affricate
98	97.97	**98.3**	Sibilant fricative
80	96.79	**100**	Silence
97	95.03	**97.91**	Stop
97	89.63	90.09	Tense
99	98.37	98.77	Velar
84	90.86	**93.47**	Voiced
92	91.31	**92.2**	Vowel
95.5	95.13	**95.60**	Average

الفصل السادس: الخاتمة والآفاق المستقبلية

1.6. خاتمة

أصبح تعلم اللغات بمساعدة الحاسوب موضوعاً هاماً بسبب الحاجة المتزايدة لتعلم لغات جديدة، ويعد كشف وتحديد النطق الخاطئ جزءاً مكملاً لأنظمة تعلم اللغات. يجري كشف النطق الخاطئ على عدة مستويات ويعد مستوى الصوتيم الأكثر دقة وفائدة لمتعلم اللغة مما دفع الباحثين للتركيز في العمل على هذا المستوى. اعتمدت منهجيات كشف النطق الخاطئ على مستوى الصوتيم على وجود مجموعات معطيات موسومة للنطق الصحيح والخاطئ لعدة لغات (الإنجليزية، الصينية، ...)، وتم بناء معظمها باستخدام شبكات التعلم العميق نظراً لفعالية النماذج الناتجة بوجود مجموعات المعطيات الصوتية الكبيرة للتدريب. استندت أنظمة تعرف الصوت لوقت طويل على سمات MFCC كونها تحاكي نظام السمع البشري وأظهرت نتائج جيدة في التطبيقات المستخدمة فيها، إلا أن كشف النطق الخاطئ يحتاج لسمات صوتية أخرى قادرة على كشف الخطأ وتحديد نوعه بدقة، لذا توجهت المنهجيات الحديثة لدراسة السمات الصوتية المتعلقة بمخارج الحروف وصفاتها وكيفية توظيفها بالشكل المناسب لتحقيق هذه المهمة. افترضنا في هذا البحث وجود نص معروف مسبقاً في نظام التعلم يقوم متعلم اللغة بقراءته، بعد ذلك تجري عملية كشف خطأ النطق. قمنا بمعالجة إشارة الكلام واستخلاص السمات الرقمية منها ثم الاستفادة من هذه السمات في مرحلتين، الأولى يتم فيها التعرف على الصوتيم باستخدام شبكة التعلم العميق LSTM وذلك للتحقق من لفظ الصوتيم نفسه، في حال عدم التعرف على الصوتيم فهذا مؤشر على وجود خطأ في النطق، وبذلك ننتقل لمرحلة ثانية لتحليل الصوت وفق مخارج وصفات الحروف كون هذا الواصفات محددة وثابتة لكل صوت. يجري في هذه المرحلة كشف الواصفات التي يحققها المتعلم عند النطق. وقد اقترحنا كشف الواصفات وفق شبكة تعلم عميق عن طريق التعلم المتعدد المهام وحقق النظام المقترح نسبة دقة أعلى من الأنظمة التي جرت المقارنة معها على مجموعتي معطيات KAPD و TIMIT. حيث حصلنا بالنسبة للمجموعة KAPD على نسبة وسطية لمقياس F1 لتعرَف الواصفات أعلى بما يقارب 4% مقارنة مع نظام آخر. وبالمثل حصلنا على أداء أفضل بالنسبة لمجموعة معطيات TIMIT. إن النتائج الجيدة لتعرف واصفات الكلام تساعد في تحديد خطأ النطق بدقة أعلى مما يجعل عملية تعلم اللغة أكثر فائدة وكفاءة. كما يمكن لهذا النموذج أن يكون جزءاً في نظام تحديد الأداء اللغوي الصوتي مثل SpeechRater المستخدم في اختبارات TOEFL.

2.6. الآفاق المستقبلية

لا يزال العمل في مجال كشف خطأ النطق وتصحيحه بشكل آلي قيد البحث والتطوير مع مختلف اللغات، إلا أن اللغة العربية لقيت قدراً أقل من العمل بسبب محدودية مجموعات المعطيات التي يمكن اعتمادها في تطوير مثل هذه الأنظمة على عكس باقي اللغات، إذ تتطلب النماذج المقترحة كمية كافية من البيانات الموسومة ليكون النظام الناتج فعالاً وقابلاً للتطبيق. يمكن التغلب على مشكلة محدودية البيانات الموسومة باستخدام شبكات الربط الزمني Connectionist temporal classification (CTC) [70]، يظهر البحث السابق فعالية شبكات CTC، فهي تلغي الحاجة لعملية التقطيع والمحاذاة القصرية التي تجري عادة للحصول على بيانات موسومة للتدريب، إضافة إلى عدم اعتمادها على وجود نماذج لغوية مدربة. على العموم يمكن توظيف نموذجنا المقترح في تطبيقات عديدة منها:

- تعليم لغة ثانوية.
- تصحيح النطق عند الأشخاص الذين يعانون من مشاكل في اللفظ.
- تقييم الأداء الصوتي للغة في أنظمة الاختبارات اللغوية.

بالنسبة لتحسين النظام المقترح يمكن تحسين مرحلة تعرف الصوتيمات من خلال:

- استكشاف سمات صوتية إضافية تزيد من قدرة الشبكة على التمييز بين الصوتيمات.
- اختبار نوع آخر من شبكات التعلم العميق يعطي دقة أعلى وأداءً أفضل.

أما مرحلة تعرف واصفات الكلام نسعى في الأبحاث القادمة لدراسة هذه الواصفات بشكل أعمق أكثر وإيجاد طريقة تكشف وتظهر خطأ النطق بشكل دقيق بالاعتماد على هذه الواصفات لتحقيق فائدة أكبر لمتعلم اللغة، فمثلاً نسعى لتجريب تدريب الواصفات التي يتلازم حدوثها معاً في الصوتيمات ضمن شبكة تعلم متعددة مهام واحدة لاختبار تحسن أداء التعرف عندها. مثل الواصفات (مهموس / احتكاكي / صامت / مستمر) نجد بينها عدة صوتيمات مشتركة، وبالتالي قد يؤدي تدريب هذا الواصفات معاً لأداء أفضل. إضافة لدراسة تأثير استخدام واصفات النطق في تحسين تعرف الصوتيمات.

المراجع

[1] N. F. Chen and H. Li, "Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*, 2017, pp. 1–7, doi: 10.1109/APSIPA.2016.7820782.

[2] K. Li, X. Qian, and H. Meng, "Mispronunciation Detection and Diagnosis in L2 English Speech Using Multidistribution Deep Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 1, pp. 193–207, Jan. 2017, doi: 10.1109/TASLP.2016.2621675.

[3] W. Li, S. M. Siniscalchi, N. F. Chen, and C. H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2016-May, pp. 6135–6139, May 2016, doi: 10.1109/ICASSP.2016.7472856.

[4] F. Nazir, M. N. Majeed, M. A. Ghazanfar, and M. Maqsood, "Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for Arabic phonemes," *IEEE Access*, vol. 7, pp. 52589–52608, 2019, doi: 10.1109/ACCESS.2019.2912648.

[5] H. Strik, K. Truong, F. de Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Commun.*, vol. 51, no. 10, pp. 845–852, 2009, doi: 10.1016/j.specom.2009.05.007.

[6] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Commun.*, vol. 67, pp. 154–166, 2015, doi: 10.1016/j.specom.2014.12.008.

[7] L. Deng, "Deep learning: from speech recognition to language and multimodal processing," *APSIPA Trans. Signal Inf. Process.*, vol. 5, pp. 1–15, Jan. 2016, doi: 10.1017/ATSIP.2015.22.

[8] M. Alghamdi, *Arabic Phonetics and Phonology*. Riyadh: Altawbah Bookshop, 2015.

[9] Noory, M. jawad. (2007). Arabic phonology. Al-Quds Open University.

[10] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. USA: Prentice Hall PTR, 2001.

[11] H. Morsy, M. Shahin, N. Aljohani, M. Shoman, and S. Abdou, "Automatic Speech Attribute Detection of Arabic Language," *Int. J. Appl. Eng. Res.*, vol. 13, no. 8, pp. 5633–5639, 2018.

[12] F. J. Owens, *Signal Processing of Speech*. London: Macmillan Education UK, 1993.

[13] D. Jurafsky and J. Martin, "Speech and Language Processing," in *Speech and Language Processing.*, vol. 3, 2014, pp. 441–458.

[14] J. Benesty, M. M. Sondhi, Y. Huang, and S. Greenberg, "Springer Handbook of Speech Processing," *J. Acoust. Soc. Am.*, vol. 126, no. 4, p. 2130, 2009, doi:

10.1121/1.3203918.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[16] I. G. and Y. B. and A. Courville, *Deep Learning*. MIT Press, 2016.

[17] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into Deep Learning," *arXiv Prepr. arXiv2106.11342*, 2021.

[18] L. Deng and D. Yu, *Automatic speech recognition*. 2015.

[19] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *J. Res. Natl. Bur. Stand. (1934).*, vol. 49, no. 6, 1952, doi: 10.6028/jres.049.044.

[20] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1–3, 1989, doi: 10.1007/BF01589116.

[21] L. Bottou, "Online Learning and Stochastic Approximations," 1998.

[22] L. Deng and D. Yu, "Deep Learning: Methods and Applications," May 2014.

[23] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Sep. 2019.

[24] Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019, doi: 10.1162/neco_a_01199.

[25] J. C. Pennington, Martha C and Richards, "Pronunciation revisited," *Teach. English to Speak. Other Lang. Inc.*, vol. 20, no. 2, pp. 207–225, 1986.

[26] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *Proc. Int. Symp. Autom. Detect. Errors Pronunciation Train. (IS ADEPT)*, no. June, pp. 1–8, 2012.

[27] M. Eskenazi, "Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype," *Lang. Learn. Technol.*, vol. 2, no. 2, pp. 62–76, 1999.

[28] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[29] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2–3, pp. 95–108, 2000.

[30] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (call)," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 1886–1890, 2013.

[31] S. Wei, G. Hu, Y. Hu, and R. H. Wang, "A new method for mispronunciation detection using Support Vector Machine based on Pronunciation Space Models," *Speech Commun.*, vol. 51, no. 10, pp. 896–905, 2009, doi: 10.1016/j.specom.2009.03.004.

[32] S. Mao, Z. Wu, X. Li, R. Li, X. Wu, and H. Meng, "Integrating Articulatory Features

into Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2018-July, 2018, doi: 10.1109/ICME.2018.8486462.

[33] A. M. Harrison, W. Y. Lau, H. M. Meng, and L. Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[34] A. Ito, Y.-L. Lim, M. Suzuki, and S. Makino, "Pronunciation error detection method based on error rule clustering using a decision tree," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[35] W.-K. Lo, S. Zhang, and H. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in *Eleventh annual conference of the international speech communication association*, 2010.

[36] M. Shahin, B. Ahmed, J. McKechnie, K. Ballard, and R. Gutierrez-Osuna, "A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[37] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2012, pp. 5049–5052.

[38] H. Strik, K. P. Truong, F. de Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[39] M. Maqsood, H. Adnan Habib, T. Nawaz, and K. Zeeshan Haider, "A Complete Mispronunciation Detection System for Arabic Phonemes using SVM," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 16, no. 3, p. 30, 2016.

[40] W. Hu, Y. Qian, and F. K. Soong, "A new Neural Network based logistic regression classifier for improving mispronunciation detection of L2 language learners," *Proc. 9th Int. Symp. Chinese Spok. Lang. Process. ISCSLP 2014*, pp. 245–249, 2014, doi: 10.1109/ISCSLP.2014.6936712.

[41] F. Hussain, M. Ehatisham-ul-haq, N. K. Baloch, and F. Ishmanov, "Improving Mispronunciation Detection of Arabic Words for Non-Native Learners Using Deep Convolutional Neural Network Features," no. June, 2020, doi: 10.3390/electronics9060963.

[42] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Transfer Learning based Non-native Acoustic Modeling for Pronunciation Error Detection," no. August, pp. 42–46, 2017.

[43] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Articulatory modeling for pronunciation error detection without non-native training data based on DNN transfer learning," *IEICE Trans. Inf. Syst.*, vol. E100D, no. 9, pp. 2174–2182, 2017, doi: 10.1587/transinf.2017EDP7019.

[44] M. Shahin and B. Ahmed, "Anomaly detection based pronunciation verification approach using speech attribute features," *Speech Commun.*, vol. 111, no. April, pp.

29–43, 2019, doi: 10.1016/j.specom.2019.06.003.

[45] W. Li, N. F. Chen, S. M. Siniscalchi, and C. H. Lee, "Improving mispronunciation detection for non-native learners with multisource information and LSTM-based deep models," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, pp. 2759–2763, 2017, doi: 10.21437/Interspeech.2017-464.

[46] M. Maqsood, A. Habib, and T. Nawaz, "An efficient mispronunciation detection system using discriminative acoustic phonetic features for Arabic consonants," *Int. Arab J. Inf. Technol.*, vol. 16, no. 2, pp. 242–250, 2019.

[47] I. Karaulov and D. Tkanov, "Attention model for articulatory features detection," Jul. 2019.

[48] M. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman, and M. A. Mekhtiche, "Towards Deep Object Detection Techniques for Phoneme Recognition," *IEEE Access*, vol. 8, pp. 54663–54680, 2020, doi: 10.1109/ACCESS.2020.2980452.

[49] M. Algabri, H. Mathkour, M. M. Alsulaiman, and M. A. Bencherif, "Deep learning-based detection of articulatory features in arabic and english speech," *Sensors (Switzerland)*, vol. 21, no. 4, pp. 1–23, Feb. 2021, doi: 10.3390/s21041205.

[50] Y. Seddiq, A. Meftah, M. Alghamdi, and Y. Alotaibi, "Reintroducing KAPD as a Dataset for Machine Learning and Data Mining Applications," in *2016 European Modelling Symposium (EMS)*, 2016, pp. 70–74, doi: 10.1109/EMS.2016.022.

[51] L. Lu, H. J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimed. Syst.*, vol. 8, no. 6, pp. 482–492, 2003, doi: 10.1007/s00530-002-0065-0.

[52] X. Lu, S. Li, and M. Fujimoto, "Automatic Speech Recognition," in *SpringerBriefs in Computer Science*, 2020, pp. 21–38.

[53] Y. Seddiq, Y. A. Alotaibi, S. A. Selouani, and A. H. Meftah, "Distinctive phonetic features modeling and extraction using deep neural networks," *IEEE Access*, vol. 7, pp. 81382–81396, 2019, doi: 10.1109/ACCESS.2019.2924014.

[54] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," Feb. 2014.

[55] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4169–4172, doi: 10.1109/ICASSP.2012.6288837.

[56] M. Learning, P. Liu, X. Qiu, and X. Huang, "Recurrent Neural Network for Text Classification," 2011.

[57] L. Chao, "Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition," pp. 65–72, 2015.

[58] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks∗," *arXiv*, 2017.

[59] N. Halabi, "Modern Standard Arabic Phonetics for Speech Synthesis," School of Electronics and Computer Science, 2016.

[60] I. Zangar *et al.*, “Duration modeling using DNN for Arabic speech synthesis To cite this version : HAL Id : hal-01889917 Duration modeling using DNN for Arabic speech synthesis,” 2018.

[61] F. K. Fahmy, M. I. Khalil, and H. M. Abbas, “A Transfer Learning End-to-End Arabic Text-To-Speech (TTS) Deep Architecture,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12294 LNAI, 2020, pp. 266–277.

[62] L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1,” *NASA STI/Recon Tech. Rep. N*, 1993, doi: 1993STIN...9327403G.

[63] K. F. Lee and H. W. Hon, “Speaker-Independent Phone Recognition Using Hidden Markov Models,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11. pp. 1641–1648, 1989, doi: 10.1109/29.46546.

[64] M. M. Alghmadi, “KACST Arabic Phonetics Database,” 2003.

[65] C. Lopes and F. Perdigao, “Phoneme Recognition on the TIMIT Database,” in *Speech Technologies*, vol. 1, InTech, 2011, pp. 285–302.

[66] M. Ravanelli, T. Parcollet, and Y. Bengio, “The PyTorch-Kaldi Speech Recognition Toolkit,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 6465–6469, Nov. 2018, doi: 10.1109/ICASSP.2019.8683713.

[67] O. Abdel-hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional Neural Networks for Speech Recognition,” vol. 22, no. 10, pp. 1533–1545, 2014.

[68] T. Bhowmik and S. K. Das Mandal, “Manner of articulation based Bengali phoneme classification,” *Int. J. Speech Technol.*, vol. 21, no. 2, pp. 233–250, 2018, doi: 10.1007/s10772-018-9498-5.

[69] M. Smit, “Cascade Deep Neural Networks Classifiers for Phonemes Recognition,” vol. 15, no. 7, pp. 1664–1670, 2020.

[70] L. Zhang, Z. Zhao, C. Ma, L. Shan, H. Sun, and L. Jiang, “End-to-End Automatic Pronunciation Error Detection Based on Improved Hybrid CTC / Attention Architecture,” 2020.